 etc. Special attention should be paid to the operation of the installation. Any handling of the product is a dangerous operation on account of the risk of error. The human factor, despite technical progress, remains a decisive element in the uncertainty of risk calculations; there is an increasing tendency to reduce the human element by automation, and to create intrinsically safe systems which can "excuse" an operator error, and even restore normal operation in the case of an incident. Despite this, humans continue to be indispensable, which is why information and workforce training are necessary to reduce risk. Active participation of all those involved in designing safety measures should be encouraged since this is the best way of achieving a safety culture in the enterprise where people are no longer the weak link in the chain, but become the central positive element of the safety system.

In most cases of manufacture or storage of dangerous substances, the consequences of a serious accident may extent far beyond the confines of the enterprise. Its siting is therefore of primary importance, the ideal being to locate it

as far as possible from populated areas. Care should also be taken to ensure that, in the case of toxic gases, the prevailing winds are not in the direction of such areas, and that, in the event of accidental spillage, the impermeability of the ground prevents pollution of underground water.

The location of dangerous sites in relation to built-up areas often presents a problem. In most European countries, demographic trends in the last two decades have led to an often uncontrolled expansion of towns and cities, which is why previously well-isolated industrial installations are now within urban boundaries, with a frightening increase in risk. Even where new enterprises are sited in rural areas, it is sometimes found that public bodies responsible for land management do little to resist pressure from landowners who wish to divide into plots and sell land that has been developed for industrial purposes, sometimes to allow company employees to live near their place of work. It is very important that any enterprise involved in the manufacture or storage of dangerous materials be enclosed by an adequate safety perimeter, including the zone where the presence of houses and dwellings presents a serious risk in the event of an accident. The safety measures adopted by the company should include alerting and informing the population involved as well as the public authorities.

This is also prescribed in European Directive 82-501 (the Seveso Directive). This covers industrial establishments (other than nuclear, military, and waste disposal) that produce, use, or store amounts of dangerous materials above the levels specified in a list appended to the directive. The directive makes the manufacturer responsible for determining the risk, taking appropriate safety measures, and informing and training the workforce. The manufacturer should also notify the relevant authorities and keep an up-to-date list of the type and amount of dangerous materials being produced and stored; describe the installation and production processes; internal emergency procedures; and provide all the necessary information for the formulation of an external emergency plan by the official safety bodies. The manufacturer should notify the authorities of all major accidents, including the circumstances, materials involved, measures adopted and to be adopted; notification has to be followed by the setting up of a special enquiry.

The safety principles governing dangerous installations in Europe are fairly well defined and regulated. An important feature of these measures is that they require regular safety checks to be carried out by an independent outside body.

Any safety system, however well designed, will tend to deteriorate over time. Not only must the safety equipment and systems be regularly checked, serviced, and repaired, but instructions to the workforce and their level of awareness and training also need to be periodically updated. There are two periods when the risk associated with an installation is greatest: during start-up, since the safety systems have not yet been tested; and after a long incident-free period, when the risk has been minimized or even forgotten by managers and workers. This is why it is essential for a disinterested observer, if possible outside the company or at least completely independent of the production organization, to make periodic visits to evaluate the safety system and update the objectives, without however interfering in the choice of measures to be adopted to achieve these objectives, which should remain the responsibility of the management.

2.4. Security

Safety may be defined as the complete absence of risk. Zero risk is an objective which one must strive for, but which is never reached. In many languages, safety and security are virtually synonymous.

Economic imperatives must be taken into account in this constant striving to improve safety. Safety has its price. No reliable safety exists if the means, and thus the necessary funding, have not been allocated for this purpose. This implies that it should not be left unsupervised in the hands of those responsible for production.

Contrary to a widespread misconception, safety cannot be achieved regardless of cost. It is therefore unacceptable that safety regulations be drawn up by a remote bureaucratic system. Even in high-level international bodies, such a system tends to add new constraints without taking into account the probability of the risk that these constraints are designed to combat, their economic effects, or sometimes even their feasibility. It is possible to arrive at provisions that

are so complex and difficult to apply that they are not implemented.

The safety costs involved in the production, transport, and storage of a dangerous product cannot exceed a certain fraction of the internationally accepted value of the product. On the other hand, the industrial safety field is so complex that there are always several ways of achieving the same result. Accordingly, the aim of experts in this field, whether they be regulatory bodies or industrial managers, should be to avoid implementing measures indiscriminately, even if they are ostensibly aimed at security, but instead choose the measures that have the best cost/benefit ratio.

Industrial safety, whether of persons or the environment, has become a political football for decision-makers and an economic debate among manufacturers. There may thus be a temptation for the former to emphasize the "public relations" impact to the detriment of efficiency, and for the latter to promote eye- catching features in order to maximize profit, all this often being supported by the media, which tend to seek sensationalism rather than objective information. Sometimes, under the pretext of safety, measures may be accepted that would otherwise have been rejected, or decisions are made purely on economic grounds.

Although these comments may appear rather pragmatic, it should not be forgotten that safety research is eminently worthwhile, even if it demands a great deal of perseverance, attention to detail, and humility. One should never believe the experts when they say, "We have done everything we can," one must continue to question. Safety techniques improve at least as quickly as production techniques, which, in the end, is very reassuring.

3. Dangerous Goods and Substances

Test methods, classification, labeling, and assigning to packaging groups of dangerous goods should be harmonized worldwide for all modes of transport (sea, rail, road, and air).

The UN Recommendations cover all modes of transport. Where less stringent requirements can be applied to a particular mode, this is indicated only in special circumstances. For air transport, more stringent requirements may occasionally apply. The Recommendations are internationally implemented through regulations and agreements covering rail, road, sea, and air transport. The parts of the Recommendations which are relevant to this chapter are the definitions of classes, principles and criteria for classification, test methods, labeling, and the *List of Dangerous Goods Most Commonly Carried* (approximately 3000 entries). There is a great degree of flexibility in the data on which the classification is based, because the ultimate classification, i.e., designation of class and inclusion in the UN Recommendations list of substances (Chap. 2) is done by UNCETDG [9] which passes judgement on the data and the proposals for classification submitted to it. The UN system uses the notion of primary hazard, which determines the hazard class. This is reflected in the recommendation that "in general not more than one danger class label should be affixed to a package". If a substance meets the definition of more than one class, a rule of precedence applies, and a distinction is made between primary and subsidiary risks (see Sections 3.1 and 3.1.10).

Principles Underlying the Regulation of the Transport of Dangerous Goods. Transport of dangerous goods is regulated in order to prevent, as far as possible, accidents to persons or property and damage to the means of transport employed or to other goods. At the same time, regulations should be framed so as not to impede the movement of such goods, other than those too dangerous to be accepted for transport. With this exception, the aim of regulations is to make transport feasible by eliminating risks or reducing them to a minimum.

Classification and Definitions of Classes of Dangerous Goods. The classification of goods by type of risk involved has been drawn up to meet technical conditions while at the same time minimizing interference with existing regulations. Classifications are made on consideration of data submitted to the Committee by governments, intergovernmental organizations, and other international organizations.

Classification of Dangerous Wastes. Wastes should be transported under the requirements of the appropriate class, considering

their hazards and the criteria of the Recommendations. Wastes not otherwise subject to the Recommendations, but covered under the Basel Convention [10] may be transported under Class 9.

Many of the substances listed in Classes 1 – 9 are deemed, without additional labeling, to be environmentally hazardous.

In Germany, the classification of "dangerous wastes", "special wastes", "dangerous residuals", "oils from wastes", and "chemical wastes" can be made on the basis of the *Abfall-Transport-Berater*, which adopts a general view on the character of the wastes/residual substances and their classification in the ADR/RID (GGVS/GGVE), taking into account the *Abfallartenkatalog*. It enables the sender, to classify such substances without too much effort [11].

Determination of the Physical State. Unless there is an explicit or implicit indication to the contrary in the UN Recommendations, dangerous goods with a melting point or initial melting point of 20 °C or lower at a pressure of 101.3 kPa are considered to be liquids. A viscous substance for which a specific melting point cannot be determined should be subjected to the ASTM D 4359-90 test; or the test for determining fluidity (penetrometer test) prescribed in Appendix A.3 of Annex A of the *European Agreement Concerning the International Carriage of Dangerous Goods by Road* (ADR) with the modifications that the penetrometer should conform to ISO 2137: 1985, and that the test should be used for viscous substances of any class.

Goods too Dangerous to be Carried Without Special Restrictions. Inherent instability of goods may involve different hazards, e.g., explosion, polymerization with intense evolution of heat, or emission of toxic gases. In most substances, such tendencies can be controlled by correct packing, dilution, stabilization, addition of an inhibitor, refrigeration, or other precautions.

Where (bearing in mind the manner in which such items are generally packed) precautionary measures are laid down for a given substance or article (e.g., that it should be "stabilized" or "inhibited") such substance or article should not normally be carried when these measures have not been taken.

Collective Entries. National or international regulations may list, either as single or as suitable collective entries, substances or articles which do not appear in the Recommendations. A "generic" or "not otherwise specified" entry (N.O.S. entry) may be used to permit the transport of substances or articles which do not appear specifically by name in the list of dangerous goods. Such a substance or article may be transported only after its dangerous properties have been determined. Any substance or article having or suspected of having explosive characteristics should first be considered for inclusion in Class 1. Some collective entries may be of the generic or N.O.S. type, provided the regulations contain provisions ensuring safety, both by excluding extremely dangerous goods from normal transport, and by covering all subsidiary risks inherent in some goods.

Classification of Solutions and Mixtures. A mixture or solution containing a dangerous substance identified by name in the Recommendations and one or more substances not subject to the Recommendations should be treated according to the requirements given for the dangerous substance, provided the packaging is appropriate to the physical state of the mixture or solution, unless:

1) The mixture or solution is specifically identified by name in the Recommendations; or
2) The entry in the Recommendations indicates that it applies only to the pure substance; or
3) The hazard class, physical state, or packing group of the solution or mixture is different from that of the dangerous substance; or
4) There is significant change in the measures to be taken in emergencies.

For a solution or mixture of which the hazard class, physical state, or packing group differs from that of the listed substance, the appropriate N.O.S. entry should be used, including its packaging and labeling provisions.

3.1. Transport Classification of Dangerous Goods

UN Classes and Divisions are summarized as follows:

Class	1	Explosives
	1.1	Substances and articles which have a mass explosion hazard
	1.2	Substances and articles which have a projection hazard, but not a mass explosion hazard
	1.3	Substances and articles which have a fire hazard and either a minor blast hazard or a minor projection hazard, or both, but not a mass explosion hazard
	1.4	Substances and articles which present no significant hazard
	1.5	Very insensitive substances which have a mass explosion hazard
	1.6	Extremely insensitive articles which do not have a mass explosion hazard
Class	2	Gases: compressed, liquefied, dissolved under pressure, or refrigerated
Class	3	Flammable liquids
Class	4.1	Flammable solids: readily combustible solids, self-reactive substances, wetted explosives
Class	4.2	Substances liable to spontaneous combustion: pyrophoric substances, self-heating substances
Class	4.3	Substances, which in contact with water, emit flammable gases
Class	5.1	Oxidizing substances
Class	5.2	Organic peroxides
Class	6.1	Toxic substances
Class	6.2	Infectious substances
Class	7	Radioactive material
Class	8	Corrosives
Class	9	Miscellaneous dangerous substances

General Provisions of ADR/RID. Annex A of ADR or Annex I of RID specifies the dangerous goods to be excluded from international carriage by road and the dangerous goods to be accepted for such carriage under certain conditions. It groups the dangerous goods in restrictive and nonrestrictive classes. Of the dangerous goods covered by the restrictive classes (Classes 1, 2, and 7) those which are listed in the clauses concerning these classes are to be accepted for carriage only under the conditions specified; others are excluded from carriage. Some of the dangerous goods covered by the nonrestrictive classes (Classes 3, 4.1, 4.2, 4.3, 5.1, 5.2, 6.1, 6.2, 8, and 9) are excluded from carriage; of the other goods covered by the nonrestrictive classes, those which are mentioned in the clauses concerning these classes are to be accepted for carriage only under the conditions specified; those not mentioned or covered by one of the collective headings are not deemed to be dangerous goods and are to be accepted for carriage without any special conditions.

ADR/RID Classes and Divisions. Hazard categories and terms used in the Regulations concerning the International Carriage of Dangerous Goods by Road and Rail (RID/ADR) [12] are summarized below.

Class 1	Explosive substances and articles	Restrictive
Class 2	Gases: compressed, liquefied, or dissolved under pressure	Restrictive
Class 3	Flammable liquids	Nonrestrictive
Class 4.1	Flammable solids	Nonrestrictive
Class 4.2	Substances liable to spontaneous combustion	Nonrestrictive
Class 4.3	Substances which, in contact with water, emit flammable gases	Nonrestrictive
Class 5.1	Oxidizing substances	Nonrestrictive
Class 5.2	Organic peroxides	Nonrestrictive
Class 6.1	Toxic substances	Nonrestrictive
Class 6.2	Infectious substances	Nonrestrictive
Class 7	Radioactive material	Restrictive
Class 8	Corrosive substances	Nonrestrictive
Class 9	Miscellaneous dangerous substances and articles	Nonrestrictive

From 1 January 1995, the classification criteria concerning flammable liquids are the same worldwide; the upper flash point limit is $\leq 61\ ^\circ C$ c.c. (closed cup test).

Hazard Categories and Terms used in the International Maritime Dangerous Goods Code (IMDG-Code) [13]. Regulation 2 of part A of Chapter VII of the *International Convention for the Safety of Life at Sea, 1974*, as amended, sets out the various classes of dangerous goods. For the purposes of the Code, it has been found necessary to subdivide a number of these classes and to define in greater detail the characteristics and properties of the substances, materials, and articles which fall within each class or division. In accordance with the criteria for selection or marine pollutants for the purposes of Annex III of the *International Convention for the Prevention of Pollution from Ships, 1973*, as modified by the Protocol of 1978 (MARPOL 73/78), a number of dangerous substances in the various classes have also been identified as substances harmful to the marine environment (Marine Pollutants, see 3.1.11).

IMO Classes and Divisions. For the following classes, the IMDG-Code differs from the UN Recommendations and the RID/ ADR:

Class 3. Flammable liquids are defined as liquids, mixtures of liquids, or liquids containing solids in solution or suspension (e.g., paints, varnishes, lacquers, but not including substances which, on account of other dangerous characteristics, have been included in other classes), which give off a flammable vapor at or below 61 °C (141 °F) closed cup (c.c.) test (corresponding to 65.6 °C (150 °F) open cup test).

In the IMDG-Code, Class 3 is further subdivided:

Class 3.1. Low-flash-point liquids; flash point < -18 °C (0 °F) c.c.

Class 3.2. Intermediate-flash point liquids; flash point from -18 °C (0 °F) up to, but not including, 23 °C (73 °F) c.c.

Class 3.3. High-flash-point liquids; flash-point from 23 °C (73 °F) up to, and including, 61 °C (141 °F) c.c.

Substances with flash point $>$ 61 °C (141 °F) c.c., are not considered dangerous by virtue of their fire hazard, where the flash point is indicated for a volatile liquid it may be followed by the symbol "c.c.", representing determination by a closed cup test, or by "o.c.", representing an open cup test. A reference to these tests is given in Section 6 of the General Introduction of the IMDG-Code.

3.1.1. Explosive Substances and Articles (Class 1)

UN Recommendations. Class 1 comprises:

1) Explosive substances (a substance which is not itself an explosive, but which can form an explosive atmosphere of gas, vapor, or dust is not included in Class 1), except those that are too dangerous to transport, or those where the predominant hazard is appropriate to another class;
2) Explosive articles, except devices containing explosive substances in such quantity of or such a character that their inadvertent or accidental ignition or initiation during transport causes no effect external to the device either by projection, fire, smoke, heat, or loud noise; and
3) Substances and articles not mentioned under (1) and (2) which are manufactured with a view to producing a practical, explosive, or pyrotechnic effect.

Transport of explosive substances which are unduly sensitive or so reactive as to be subject to spontaneous reaction is prohibited.

Under the UN Recommendations, the following definitions apply:

1) An explosive substance is a solid or a liquid substance (or a mixture of substances) which is in itself capable by chemical reaction of producing gas at such a temperature and pressure and at such a speed as to cause damage to the surroundings. Pyrotechnic substances are included even when they do not evolve gases.
2) A pyrotechnic substance is a substance or mixture of substances designed to produce an effect by heat, light, sound, gas, or smoke, or a combination of these as the result of nondetonative self-sustaining exothermic chemical reactions.
3) An explosive article is an article containing one or more explosive substances.

Class 1 is divided into six divisions and 13 compatibility groups.

Division 1.1. Substances and articles which have a mass explosion hazard (a mass explosion is one which affects almost the entire load virtually instantaneously).

Division 1.2. Substances and articles which have a projection hazard, but not a mass explosion hazard.

Division 1.3. Substances and articles which have a fire hazard and either a minor blast hazard or a minor projection hazard or both, but not a mass explosion hazard. This division comprises substances and articles: which give rise to considerable radiant heat; or which burn one after another, producing minor blast or projection effects, or both.

Division 1.4. Substances and articles which present no significant hazard. This division comprises substances and articles which present only a small hazard in the event of ignition or initiation during transport. The effects are largely confined to the package and no projection of fragments of appreciable size or range is to be expected. An external fire should not cause virtu-

ally instantaneous explosion of almost the entire contents of the package.

Division 1.5. Very insensitive substances which have a mass explosion hazard. This division comprises substances which have a mass hazard, but are so insensitive that there is very little probability if initiation or of transition from burning to detonation under normal conditions of transport. The probability of transition from burning to detonation is greater when large quantities are carried in a ship.

Division 1.6. Extremely insensitive articles which do not have a mass explosion hazard. This division comprises articles which contain only extremely insensitive detonating substances, and which demonstrate a negligible probability of accidental initiation or propagation; the risk is limited to the explosion of a single article.

Class 1 is unique in that the type of packaging frequently has a decisive effect on the hazard, and therefore on the assignment to a particular division. The correct division and compatibility group and the packaging are determined by the method outlined in Chapter 4 of the UN Recommendations.

The definition and the classification in divisions and compatibility groups of the UN Recommendations corresponded the those of IMO and RID/ADR.

3.1.2. Gases (Class 2)

UN Recommendations. A gas is a substance which:

1) At 50 °C has a vapor pressure > 300 kPa;
2) Is completely gaseous at 20 °C at a standard pressure of 101.3 kPa.

The transport condition of a gas is described according to its physical state as:

1) Compressed gas: a gas (other than in solution) which, when packaged under pressure for transport, is entirely gaseous at 20 °C;
2) Liquefied gas: a gas which when packaged for transport is partially liquid at 20 °C;
3) Refrigerated liquefied gas: a gas which when packaged for transport is partially liquid because of its low temperature; or
4) Gas in solution: compressed gas which when packaged for transport is dissolved in a solvent.

The class includes compressed gases, liquefied gases, gases in solution, refrigerated liquefied gases; mixtures of gases, mixtures of one or more gases with one or more vapors of substances from other classes, articles charged with a gas, tellurium hexafluoride, and aerosols.

Class 2 substances are assigned to one of three divisions based on the primary hazard during transport.

Division 2.1. Flammable gases. Gases which at 20 °C and a standard pressure of 10.3 kPa:

1) Are ignitable in a mixture of 13 vol % or less with air; or
2) Have a flammable range with air of at least 12 %, regardless of the lower flammability limit. Flammability should be determined by tests or by calculation in accordance with ISO 10 156: 1990 (where insufficient data are available, tests by a comparable method recognized by a national competent authority may be used).

Division 2.2. Non flammable, nontoxic gases. Gases which are transported at a pressure not less than 280 kPa at 20 °C, or as refrigerated liquids, and which:

1) Are asphyxiant gases, which dilute or replace the oxygen normally in the atmosphere; or
2) Are oxidizing gases, which may, generally by providing oxygen, cause or contribute to the combustion of other material more than air does; or
3) Do not come under the other divisions.

Division 2.3. Toxic gases. Gases which:

1) Are known to be so toxic or corrosive to humans as to pose a hazard to health; or
2) Are presumed to be toxic or corrosive to humans, because they have $LC_{50} \leq 5000$ mL/m^3 (ppm) when tested in accordance with Chapter 6 subparagraph 6.5 (c) of the Recommendations. Gases meeting the above criteria owing to their corrosive properties are classified as toxic with a subsidiary corrosive risk.

Mixtures of Gases. For the classification of gas mixtures into one of the three divisions (including vapors of substances from other classes) the principles of Chapter 1 of the UN Recommendations can be used.

Hazard Precedence. Gases and gas mixtures with hazards associated with more than one division take the following precedence:

1) Division 2.3 takes precedence over all other divisions
2) Division 2.1 takes precedence over Division 2.2

ADR/RID Definitions and Divisions of Class 2. Substances with critical temperature $< 50\,°C$, or at $50\,°C$, a vapor pressure greater than 300 kPa (3 bar) are deemed to be Class 2 substances. Substances and articles of Class 2 are classified as follows:

A) Compressed gases with critical temperature $< -10\,°C$
B) Liquefied gases with critical temperature $\geq -10\,°C$:
   a) Liquefied gases with critical temperature $\geq 70\,°C$
   b) Liquefied gases with critical temperature $\geq -10\,°C$, but $< 70\,°C$;C.
C) Deeply refrigerated liquefied gases
D) Gases dissolved under pressure
E) Aerosol dispensers and nonrefillable containers of gas under pressure
F) Gases subject to special requirements
G) Empty receptacles and empty tanks

The substances and articles of Class 2 are subdivided according to their chemical properties, as follows:

(a) Nonflammable
(at) Nonflammable, toxic
(b) Flammable
(bt) Flammable, toxic
(c) Chemically unstable
(ct) Chemically unstable, toxic

Unless otherwise specified, chemically unstable substances are considered to be flammable.

IMO Definitions and Divisions of Class 2 are the same as those in the UN Recommendations.

3.1.3. Flammable Liquids (Class 3)

UN Recommendations. Flammable liquids are liquids, or mixtures of liquids, or liquids containing solids in solution or suspension (paints, varnishes, lacquers, etc., but not including substances otherwise classified on account of their dangerous characteristics) which give off a flammable vapor at temperatures of not more than $60.5\,°C$ c.c., or not more than $65.6\,°C$ o.c., normally referred to as the flash point. However such liquids with a flash point of more than $35\,°C$, which do not sustain combustion need not be considered as flammable liquids for the purposes of the UN Recommendations. Liquids offered for transport at temperatures at or above their flash point are in any case considered as flammable liquids. Flammable liquids also include substances that are transported or offered for transport at elevated temperatures in a liquid state, and which give off a flammable vapor at or below the maximum transport temperature.

Liquids are considered to be noncombustible if they have passed a suitable combustibility test, if their fire point according to ISO 2592: 1973 is greater than $100\,°C$, or if they are water-miscible solutions with a water content of more than 90 wt %.

Viscous substances with flash point $< 23\,°C$ may be placed in Packing Group III. Substances classified as flammable liquids owing to their being transported or offered for transport at elevated temperatures are included in Packing Group III.

ADR/RID Definitions and Divisions of Class 3.

1) Flammable liquids within the meaning of ADR, are those that are liquid at the maximum temperature of $20\,°C$, or for viscous substances for which a specific melting point cannot be determined, and which are highly viscous according to the criteria of the penetrometer test (see Appendix A.3 of ADR and Appendix III of RID), or are liquid according to the ASTM D 4359-90 test method
2) Have at $50\,°C$ a vapor pressure ≤ 300 kPa (3 bar)
3) Have a flash point of $\leq 61\,°C$ c.c.

Class 3 substances, are assigned to one of the following groups designated (a), (b), or (c) according to their degree of danger:

(a) Very dangerous substances: flammable liquids with boiling point or initial boiling point not exceeding 35 °C, and flammable liquids with flash point < 23 °C which are either highly toxic according to the criteria of Class 6.1 or highly corrosive according to the criteria of Class 8
(b) Dangerous substances: flammable liquids with flash point < 23 °C which are not classified under (a), with the exception of viscous substances
(c) Substances presenting a minor danger: flammable liquids with flash point 23–61 °C, and some viscous substances

Methods used for grouping, and combustibility testing of flammable liquids as well as methods for grouping of flammable viscous substances with flash point < 23 °C can be found in Appendix A.3 of ADR and Appendix III of RID, respectively. Viscosity and flash point data are summarized in Table 1.

Table 1. Viscosity and flash point data of Class 3 substances

Kinematic viscosity ν (extrapolated), at near-zero shear rate at 23 °C, mm^2/s	Flow time t (ISO 2431-1984), s	Jet diameter, mm	Flash point, °C
20–80	20–60	4	>17
80–135	60–100	4	>10
135–220	20–32	6	> 5
220–300	32–44	6	>−1
300–700	44–100	6	>−5
>700	>100	6	≤ −5

IMO definitions and divisions of Class 3 are described in Section 3.1.

3.1.4. Flammable Solids, Substances Liable to Spontaneous Combustion, and Substances which in Contact with Water Emit Flammable Gases (Class 4)

UN Recommendations. Class 4 comprises:

Division 4.1. Flammable solids. Solids which, under conditions encountered in transport, are readily combustible or may cause or contribute to fire through friction; self-reactive and related substances which are liable to undergo a strongly exothermic reaction; desensitized explosives which may explode if not diluted sufficiently.

Division 4.1 comprises the following types of substances:

(a) Flammable solids
(b) Self-reactive and related substances
(c) Desensitized explosives

Division 4.2. Substances liable to spontaneous combustion under conditions encountered in transport, or to heating up in contact with air, and which are then liable to catch fire;

Division 4.2 comprises:

(a) Pyrophoric substances
(b) Self-heating substances

Division 4.3. Substances which in contact with water emit flammable gases. Substances which, by interaction with water, are liable to become spontaneously flammable or to give off flammable gases in dangerous quantities.

3.1.4.1. Flammable Solids [Division 4.1 (a)]

Properties. Flammable solids are readily combustible solids and solids which may cause fire through friction. Readily combustible solids are powders or granules, ignited by brief contact with an ignition source, such as a burning match, provided the flame spreads rapidly. The danger may come not only from the fire, but also from toxic combustion products. Metal powders are especially dangerous because of the difficulty of extinguishing a fire since normal extinguishing agents such as carbon dioxide or water can increase the hazard.

Classification of flammable solids should be carried out according to Chapter 14 of the UN Recommendations.

3.1.4.2. Self-Reactive and Related Substances [Division 4.1 (b)]

Definitions. Self-reactive substances are thermally unstable substances liable to undergo a strongly exothermic decomposition, even without participation of oxygen (air). Substances should be excluded from Division 4.1, if:

1) They are explosives, according to the criteria of Class 1

2) They are oxidizing substances, according to the assignment procedure of Division 5.1
3) They are organic peroxides, according to the criteria of Division 5.2
4) Their heat of decomposition is < 300 J/g
5) Their self-accelerating decomposition temperature (SADT) is $> 75\,^{\circ}$C for a 50-kg package

Heat of decomposition can be determined by any internationally recognized method, e.g., differential scanning calorimetry, adiabatic calorimetry. Any substance which shows the properties of a self-reactive substance should be classified as such, even if this substance gives a positive test result for inclusion in Division 4.2. Substances related to self-reactive substances are distinguished from the latter by having a self-accelerating decomposition temperature $> 75\,^{\circ}$C.

Properties. The decomposition of self-reactive substances can be initiated by heat, contact with catalytic impurities (e.g., acids, heavy-metal compounds, bases), friction, or impact. The rate of decomposition increases with temperature and varies with the substance. Decomposition, particularly if no ignition occurs, may result in the evolution of toxic gases or vapors. For certain self-reactive substances, the temperature should be controlled. Some self-reactive substances may decompose explosively, particularly if confined. This characteristic may be modified by the addition of diluents or by the use of appropriate packaging. Some self-reactive substances burn vigorously.

Self-reactive substances include: aliphatic azo compounds; organic azides; diazonium salts; *N*-nitroso compounds; aromatic sulfohydrazides.

Classification of self-reactive and related substances should be carried out according to Chapter 14 of the UN Recommendations.

Temperature Control Requirements. Self-reactive substances should be subject to temperature control if their self-accelerating decomposition temperature (SADT) is $\leq 55\,^{\circ}$C. Test methods for determining the SADT are given in the current edition of the Recommendations on the Transport of Dangerous Goods, Manual of Tests and Criteria. The test selected should be conducted in a manner which is representative, both in size and material, of the package to be transported.

3.1.4.3. Desensitized Explosives [Division 4.1 (c)]

Properties. Desensitized explosives are substances which are wetted with water or alcohol or diluted with other substances to suppress their explosive properties.

3.1.4.4. Substances Liable to Spontaneous Combustion [Division 4.2]

Properties. Self-heating of substances, leading to spontaneous combustion, is caused by reaction of the substance with oxygen (in the air), conduction of the heat developed not being rapid enough to avoid combustion. Spontaneous combustion occurs when the rate of heat production exceeds the rate of the heat loss and the autoignition temperature is reached. Two types of substance can be distinguished with spontaneous combustion properties:

(a) Substances, including mixtures and solutions (liquid or solid), which even in small quantities ignite within 5 min of coming into contact with air; these substances are the most liable to spontaneous combustion and are called pyrophoric substances.
(b) Other substances which in contact with air, without addition of energy, are liable to self-heating; these substances ignite only when present in large amounts (kilograms) and after long periods of time (hours or days), and are called self-heating substances.

Test methods for pyrophoric solids and liquids as well as for self-heating substances can be found in Chapter 14 of the UN Recommendations.

3.1.4.5. Substances which in Contact with Water Emit Flammable Gases (Division 4.3)

Properties. Certain substances in contact with water may emit flammable gases that can form explosive mixtures with air. Such mixtures

are easily ignited by all ordinary sources of ignition, e.g., naked lights, electrical sparts, or unprotected light bulbs. The resulting blast wave and flames may endanger people and the environment. A suitable test should be used to determine whether the reaction of a substance with water produces a dangerous amount of gases which may be flammable; it should not be applied to pyrophoric substances.

Methods of classification and assignment to packaging groups can be found in Chapter 14 of the UN Recommendations.

3.1.5. Oxidizing Substances and Organic Peroxides (Class 5)

UN Recommendations. This class comprises:

Division 5.1. Oxidizing substances which, while in themselves are not necessarily combustible, may, generally by yielding oxygen, cause, or contribute to, the combustion of other material.

Division 5.2. Organic peroxides which contain the bivalent – O – O – group and may be considered as hydrogen peroxide derivatives, with one or both of the hydrogen atoms replaced by organic radicals. Organic peroxides are thermally unstable substances, which may undergo exothermic self-accelerating decomposition.

Methods of classification and assignment to packaging groups can be found in Chapter 11 of the UN Recommendations.

Assignment of Organic Peroxides to Division 5.2. Any organic peroxide should be considered for classification in Division 5.2, unless the formulation contains:

1) Not more than 1.0 % available oxygen from the organic peroxides, when containing not more than 1.0 % hydrogen peroxide; or
2) Not more than 0.5 % available oxygen from the organic peroxides, when containing more than 1.0 %, but not more than 7.0 % hydrogen peroxide.

Allocation of new organic peroxides, or new formulations, or mixtures of currently assigned organic peroxides to a generic entry should be made by the competent authority of the country of origin, on the basis of a test report.

Methods for classification and assignment to packaging and packaging groups of organic peroxides can be found in Chapter 11 of the UN Recommendations. Suitable test methods with pertinent evaluation criteria are given in the current edition of Recommendations on the Transport of Dangerous Goods, Manual of Tests and Criteria, Part III.

Temperature Control. To avoid unnecessary confinement, metal packagings meeting the test criteria of Packing Group I should not be used. Organic peroxides are assigned to Packing Group II (medium danger).

3.1.6. Toxic and Infectious Substances (Class 6)

UN Recommendations. This class comprises:

Division 6.1. Toxic substances are liable either to cause death or to harm human health if swallowed, or inhaled, or by skin contact.

Division 6.2. Infectious substances contain viable microorganisms, including a bacterium, virus, Rickettsia, parasite, fungus, or a recombinant, hybrid or mutant, that are known or are reasonably believed to cause disease in animals or humans.

Genetically modified microorganisms and organisms which do not meet the definition of an infectious substance should be considered for classification in Class 9 and assignment to UN 3245.

Toxins from plant, animal, or bacterial sources which do not contain infectious substances or organisms, or which are not contained in them should be considered for classification in Division 6.1 and assignment to UN 3172.

3.1.6.1. Toxic Substances (Division 6.1)

Criteria for Defining Toxicity. Substances of Division 6.1, including pesticides, should be allocated among the three packing groups according to their degree of toxic hazard in transport as follows:

Packing Group I:	substances and preparations presenting a very severe toxicity risk
Packing Group II:	substances and preparations presenting a serious toxicity risk
Packing Group III:	substances and preparations presenting a relatively low toxicity risk

In making this grouping, account should be taken of experience in instances of accidental poisoning in humans, and of special properties possessed by any individual substance, such as liquid state, high volatility, any special likelihood of penetration, and special biological effects.

In the absence of human experience, grouping should be based on data obtained from animal experiments. Three possible routes of administration should be examined: oral ingestion; dermal contact; and inhalation of dust, mist, or vapor. Appropriate animal tests for the various routes of exposure are described in Chapter 6 of the UN Recommendations. When a substance exhibits a different order of toxicity by two or more of these routes of administration, the highest degree of danger indicated by the tests should be assigned.

The criteria for grouping a substance according to the toxicity it exhibits by all three routes of administration are presented in Chapter 6 of the UN Recommendations.

3.1.6.2. Infectious Substances (Division 6.2)

Infectious substances contain viable microorganisms: bacterium, virus, *Rickettsia*, parasite, fungus, or a recombinant, hybrid or mutant, that are known or reasonably believed to cause disease in animals or humans. However, they are not subject to the recommendations for this division if the spread of disease to humans or animals exposed to such substances is considered unlikely.

Infectious substances should be considered for classification in Division 6.2 and for assignment to UN 2814 (UN-No. 3245: genetically modified microorganisms) or UN 2900 after allocation to four risk groups, based on criteria developed by the World Health Organization.

Genetically modified microorganisms and organisms are microorganisms and organisms in which genetic material has been altered by genetic engineering.

Biological products are either finished biological products for human or veterinary use, manufactured in accordance with the requirements of national public health authorities, transported, if required, under special approval or license from such authorities; or biological products transported prior to licensing for development or investigation; or finished products for experimental treatment of humans or animals, manufactured in compliance with the requirements of national public health authorities. They also cover unfinished biological products, prepared in accordance with procedures of specialized government agencies.

Some licensed vaccines may present a biohazard in certain parts of the world only. In that case, competent authorities may require these vaccines to comply with the requirements for infectious substances, or may impose other restrictions.

Diagnostic specimens are any human material including, but not limited to, excreta, secreta, blood and its components, tissue and tissue fluids, transported for diagnostic or investigation purposes, but excluding live infected animals.

Wastes transported under UN 3291 (UN-No. 3291: clinical waste, unspecified, n.o.s. or (bio) medical waste, n.o.s. or regulated medical waste, n.o.s.) are wastes derived from the medical treatment of animals or humans or from bioresearch where there is a relatively low probability that infectious substances are present. Waste infectious substances which can be specified should be assigned to UN 2814 (UN-No. 2814: infectious substances, affecting humans) or to UN 2900 (UN-No. 2900: infectious substances, affecting animals only). Decontaminated wastes which previously contained infectious substances should be considered nondangerous, unless the criteria of another class are met.

These requirements are also applicable to RID/ADR and IMO from 1 January 1995.

3.1.7. Radioactive Material (Class 7)

UN Recommendations. A radioactive material is defined as any material with specific activity > 70 kBq/kg (0.002 μCi/g). In this context, specific activity means the activity per unit mass of radionuclide or, for a material in which the radionuclide is essentially uniformly distributed, the activity per unit mass of material.

Regulations regarding the transport of radioactive material have been prepared by the International Atomic Energy Agency (IAEA) in consultation with the United Nations, specialized agencies, and IAEA Member States.

3.1.8. Corrosive Substances (Class 8)

UN Recommendations. These are substances which, by chemical action, cause severe damage on contact with living tissue, or, in the case of leakage, materially damage, or even destroy, other goods or the means of transport; they may also cause other hazards.

Special Recommendations Relating to Class 8. Substances and preparations of Class 8 are divided among three packing groups, according to their degree of hazard in transport as follows:

Packing Group I:	very dangerous substances and preparations
Packing Group II:	substances and preparations presenting medium danger
Packing Group III:	substances and preparations presenting minor danger

Allocation of substances to the packing groups in Class 8 has been on the basis of experience, taking into account such additional factors as inhalation risk and reactivity with water (including the formation of dangerous decomposition products). [A substance or preparation meeting the criteria of Class 8 having an inhalation toxicity of dusts and mists (LC_{50}) in the range of Packing Group I, but toxicity through oral ingestion or dermal contact only in the range of Packing Group III or less, should be allocated to Class 8.] New substances, including mixtures, can be judged by the contact time necessary to produce full-thickness destruction of human skin. Substances judged not to cause full-thickness destruction of human skin should still be considered for their potential to cause corrosion of metal surfaces.

In making this grouping, account should be taken on experience in instances of accidental exposures to humans. In the absence of such experience grouping should be based on data obtained from animal experiments in accordance with OECD Guideline 404 [15].

The test criteria for the three groups in this class are:

Packing Group I (very dangerous substances): Substances that cause full-thickness destruction of intact skin tissue within an observation period of 60 min, starting after an exposure time of 3 min or less.

Packing Group II (substances presenting medium danger): Substances that cause full-thickness destruction of intact skin tissue within an observation period of 14 d, starting after an exposure time of more than 3 min, but not more than 60 min.

Packing Group III (substances presenting minor danger):

(a) Substances that cause full-thickness destruction of intact skin tissue within an observation period of 14 d, starting after an exposure time of more than 60 min, but not more than 4 h.
(b) Substances which are judged not to cause full-thickness destruction of intact skin tissue, but which exhibit a corrosion rate on steel or aluminum surfaces exceeding 6.25 mm/a at 55 °C. For the purposes of testing steel, type P 3 [ISO 2604 (IV): 1975] or a similar type, and for testing aluminum, non-clad types 7075-T 6 or AZ 5 GU-T 6 should be used.

3.1.9. Miscellaneous Dangerous Substances and Articles (Class 9)

UN Recommendations. These are substances and articles which during transport present a danger not covered by other classes, including substances that are transported or offered for transport at tempertures ≥ 100 °C in a liquid state or at temperatures ≥ 240 °C in a solid state.

IMO Definitions. Class 9 comprises:

1) Substances and articles not covered by other classes which experience has shown, or may show, to be of such a dangerous character that the provisions of part A Chapter VII of the International Convention for the Safety of Life at Sea, 1974, as amended, should apply
2) Substances not subject to the provisions of part A of Chapter VII of the aforementioned Convention, but to which the regulations of Annex III of the International Convention for the Prevention of Pollution from Ships, 1973, as modified by the Protocol of 1978 (MARPOL 73/78), apply

ADR/RID Definitions. Class 9 covers substances and articles which, during carriage, present a danger not covered by the headings of other classes; it includes:

Substances which, on inhalation as fine dust, may endanger health
Substances and apparatus which in the event of fire may form dioxins
Substances evolving flammable vapor
Lithium batteries

From 1 January 1995, a new subsection "Environmentally hazardous substances" should be implemented, including UN 3077 (Environmentally hazardous substances, solid, n.o.s.), and UN 3082 (Environmentally hazardous substances, liquid, n.o.s).

These UN numbers take into account inland water pollution effects as well as genetically modified microorganisms and organisms.

3.1.10. Estimation of Major Hazards when there are Several Hazardous Properties

Precedence of Hazard Characteristics of the UN Recommendations. Table 1.44 of the UN Recommendations may be used as a guide in determining the class of a substance, mixture, or solution having more than one risk, when it is not named in the list of dangerous goods in the UN Recommendations. For goods having multiple risks which are not specifically listed, the most stringent packing group assigned to the respective hazards of the goods takes precedence over other packing groups, irrespective of the precedence of the hazard table. The precedence of hazard characteristics of the following have not been dealt with in the table, as these primary characteristics always take precedence:

Class 1 substances and articles
Class 2 gases
Division 4.1 self-reactive and related substances and desensitized explosives
Division 4.2 pyrophoric substances
Division 5.2 substances
Division 6.1 substances with a Packing Group I inhalation toxicity, except for substances or preparations meeting the criteria of Class 8 having an inhalation toxicity of dusts and mists (LC_{50}) in the range of Packing Group I, but toxicity through oral ingestion or dermal contact only in the range of Packing Group III or less, which should be allocated to Class 8
Division 6.2 substances
Class 7 material

3.1.11. Seawater Pollutants and Inland Water Pollutants

In Germany, the individual federal states are responsible for regulating the transport of inland water pollutants, as reflected in the *Wasserhaushaltsgesetz* (WHG Paragraph 19 G). From 1 January 1995 the transport of inland water pollutants is regulated by the new section F of Class 9 marginals 11 and 12 ADR/RID.

Seawater Pollutants. At the International Conference of Marine Pollutants, 1973, the need was recognized to preserve the marine environment. Provisions are contained in Annex III of the *International Convention for the Prevention of Pollution from Ships, 1973*, as modified by the Protocol of 1978 (MARPOL 73/78). The Marine Environment Protection Committee, at its 21st session in 1985, decided that Annex III should be implemented through the International Maritime Dangerous Goods Code (IMDG-Code). This decision was endorsed by the Maritime Safety Committee at its 51st session in 1985.

Substances, materials, or articles are identified in the individual schedules of the IMDG-Code as Marine Pollutants. Substances, articles

or materials that are so identified, but do not possess any other hazard, are listed in the appendices to the schedules for UN 3077, or UN 3082, in Class 9.

Solutions, Mixtures, and Isomers. A solution or mixture containing $\geq 10\,\%$ of one or more Marine Pollutants is itself a Marine Pollutant.

Certain Marine Pollutants have an extreme pollution potential and are identified as Severe Marine Pollutants in the individual schedules. A solution or mixture containing 1 % or more of these Severe Marine Pollutants is a Marine Pollutant.

A solution or mixture which does not fall within the criteria of Classes 1 – 8, but which meets the criteria for Marine Pollutants should be offered for transport as an Environmentally Hazardous Substance, Solid, N.O.S, or as an Environmentally Hazardous Substance, Liquid, N.O.S, under Class 9, even though it is not listed by name.

An isomer of a substance identified as a Marine Pollutant and covered by a generic entry in Classes 1 – 8, but which does not fall within the criteria of these classes, should be offered for transport as an Environmentally Hazardous Substance, Solid, N.O.S, or as an Environmentally Hazardous Substance, Liquid, N.O.S, under Class 9, even though it is not listed by name.

Packages containing Marine Pollutants should be durably marked with the Marine Pollutant Mark; marking should not apply to:

1) Packages containing Marine Pollutants in inner packagings with contents of: 5 L or less for liquids; or 5 kg or less for solids
2) Packages containing Severe Marine Pollutants in inner packagings with contents of: 0.5 L or less liquids; or 500 g or less for solids

In the case of loss or likely loss overboard of marked Marine Pollutants into the sea, the master or other person in charge of the ship concerned must report the particulars of such loss or likely loss overboard, wherever it occurs, by the fastest telecommunication channel available with the highest possible priority to the nearest coastal state.

Inland Water Pollutants. In 1995 new requirements for the inclusion of Environmentally Hazardous Substances (Aquatic Pollutants) came into force. The criteria are consistent with the final text for substances pollutant to the aquatic environment adopted by the EC in Commission Directive No. 91/325/EEC, Section 5 [16].

In Appendix III/A.3 of RID/ADR a new Section G has been added test methods to determine the ecotoxicity, persistance and bioaccumulation of substances in the aquatic environment for classification in Class 9. The test methods used shall be those adopted by the OECD and EC. If other methods are used, they shall be internationally recognized, be equivalent to the OECD/EC tests and be referenced in the test report.

Tests to determine aquatic pollutant effects are as follows:

Acute toxicity for fish
Acute toxicity for *Daphnia*
Algal growth inhibition
Tests for bioaccumulation potential

3.2. Storage of Dangerous Goods

In Germany, the Technical Regulations for Dangerous Substances (*Technische Regeln für Gefahrstoffe*, TRGS) established by the Committee of Dangerous Substances (*Ausschuß für Gefahrstoffe* (AGS) reflect, albeit never definitively and completely, the state of regulations and knowledge mentioned in § 17 of the Directive of Dangerous Substances (*Gefahrstoff V*), especially for handling purposes. This also involves legal considerations of the storage of dangerous goods and substances.

The regulations to be applied are as follows:

- TRGS 511: storage of ammonium nitrate
- TRGS 514: storage of severely toxic and toxic substances in packaging and in portable tanks
- TRG 515: storage of oxidizing substances in packaging and in portable tanks

3.2.1. Explosive Substances (Class 1)

The storage of explosive substances is regulated by the Second Directive to the Explosive Law

(*2. SprengV*) in its latest edition of 5 September 1989, the Technical Regulation for the Storage of Explosives, which are subject to 2. SprengV and the section on Accident Prevention Regulation (*Unfallverhütungsvorschrift*, UVV) 46 a *Explosive Substances and Articles with Explosive Substances*, General rule (VBG 55 a) established by the Employers' Liability Insurance Association of Chemical Industries (*Berufsgenossenschaft der chemischen Industrie*). In 2. SprengV in §1, item (2), No. 1 and 2, the area of application is clearly described. The Directive is not applicable to explosive substances during transportation by road, rail, water, or air, and on vehicles mentioned in § 1, item (2), No. 1 and during loading or unloading.

Storage Groups. Explosives and explosive articles are classified in four storage groups, by analogy to the transport Divisions 1.1 – 1.4. There are no storage groups corresponding to Divisions 1.5 and 1.6.

The assignment of substances and articles to storage groups is normally based on results of tests which are carried out by the Federal Institute for Materials Research and Testing (BAM). These tests and the criteria for the assignment to storage groups are described in the Rules for the Storage of Explosives (*Sprengstofflagerrichtlinie*, Spreng LR 010). Of most importance are the properties of the substances and articles in their packagings, in connection with their degree of risk to the public. Explosive substances from the chemical industry are also covered by the German Explosives Law, but for transportation purposes are normally placed in classes other than Class 1 (e.g., Classes 4.1 or 5.2). Such explosive chemicals are also assigned to certain storage groups on the basis of experimental procedures described the Spreng LR 011.

Compatibility Groups. According to certain properties of the substances or articles, a compatibility group is determined (A, B, C, D, E, F, or G). The definitions of the compatibility groups correspond to those used for transport.

3.2.2. Flammable Liquids (Class 3)

For storage of flammable liquids in Germany specific Technical Regulations (*Technische Regeln für brennbare Flüssigkeiten*, TRbF) are valid. Their objective, as part of the according Directive (*Verordnung über brennbare Flüssigkeiten*, VbF), is to regulate filling, loading, and unloading of flammable liquids.

The regulations are divided as follows:

- TRbF 001: general provisions, structure, and applicability
- TRbF 002: standards, etc.
- TRbF 003: testing of flammable liquids
- TRbF 100/200: general safety requirements
- TRbF 111/211: places of filling, emptying
- TRbF 120/220: general provisions for metallic or nonmetallic fixed tanks
- TRbF 121/221: specific provisions for metallic or nonmetallic fixed tanks

According to the VbF, flammable liquids are those which are liquid or viscous at a temperature not exceeding 35 °C and which have vapor pressure not exceeding 300 kPa (3 bar) at a temperature of 50 °C.

3.2.3. Oxidizing Substances (Class 5.1)

The storage of oxidizing substances in packaging and in portable tanks should be in accordance with the provisions of the technical rule TRGS 515. This rule is not valid for substances in process, handling, storage under the provision for the transport of dangerous goods, and amounts of substances stored up to 200 kg. From 1 October 1993, the storage and handling of organic peroxides is regulated by a new Accident Prevention Regulation (VBG 58) published by the Employer's Liability Insurance Association of Chemical Industries Peroxides are assigned, according to their danger, to storage groups, similar to those for explosive substances.

3.2.4. Toxic Substances (Class 6.1)

The storage of toxic substances in packaging and in portable tanks should be in accordance with the provisions of the technical rule TRGS 514. This rule is not valid for substances in process, handling, storage under the provision for the transport of dangerous goods, and amounts of substances stored up to 200 kg (50 kg for very toxic substances).

3.2.5. Radioactive Material (Class 7)

Within the nuclear fuel cycle, the safe storage of different types of radioactive material is necessary. An important product for nuclear fuel fabrication is uraniumhexafluoride UF_6 which is shipped and stored in steel cylinders designed to withstand a test pressure of 2.8 MPa (28 bar).

Spent nuclear fuel must be stored after removal from the reactor, first in water ponds within the nuclear power plant. This first stage of storage (when most of the decay heat is dissipated) is followed by reprocessing or interim storage for a longer period, up to 50 years. There are several interim storage options for spent fuel or high-level waste:

- Dry storage in an inert gas atmosphere in casks, silos, or vaults
- Wet storage in water ponds

In every case, the technical measures have to satisfy the following protection aims:

Shielding against radiation
Containment of the radioactive inventory
Nuclear safety
Heat removal

The back-end of the nuclear fuel cycle is the disposal of the radioactive waste products, or the direct disposal of spent nuclear fuel.

In some countries, low-level or non-heat-generating waste is disposed of in above-ground repositories (United States, France); in other countries in deep underground repositories (Sweden, Germany). High-level waste or spent fuel should be disposed of only in deep geological formations. The general safety principles for the disposal of radioactive materials are:

- Immobilization of radioactive isotopes (e.g., by vitrification, cementation)
- Containment by technical and geotechnical barriers during the service period of the repository
- Long-term containment by geological barriers

4. Safety Measures

To minimize the risk in transportation, handling, and storage of hazardous substances appropriate safety measures must:

1) Assure safe service conditions
2) Reduce the probability of incidents or accidents
3) Reduce the consequences of accidental release

Safety measures can be classified as:

1) Measures to assure safe confinement of the hazardous substance during normal service conditions, abnormal service conditions (incidents), and accidents (i.e., accident prevention measures)
2) Measures to reduce the consequences of accidental losses (i.e., accident limitation measures)

Besides this "horizontal" differentiation, a "vertical" division of safety measures into:

Technical safety measures
Operational safety measures

is possible. Table 2 shows the combination of the main components of safety measures. To ensure that all measures are met effectively and within specified limits, qualitiy assurance is necessary, at every stage.

The most interesting problem in safety science and technology is the correct choice (i.e., the selection of the appropriate combination and technical level) of measures to provide a level of safety, accepted on commercial considerations by the supplier, on safety grounds by the workers, and on environmental aspects by the public. Figure 1 shows the correlation between incident probability and consequences. If the specific hazard potential of substance 2 (case 2) is bigger than that of substance 1 (case 1), only a lower "consequence grade" (e.g., a lower release rate) in case of an accident can be accepted; the consequences for substance 2 have to be restricted to the site area, but for substance 1 an accident consequence may be accepted outside. With respect to the safety measures necessary to reduce the incident probability to an acceptable level, the minimum safety in case 2 has to be higher than in case 1.

Table 2. Combination of the main components of safety measures

Safety measures	Normal service safety measures	Accident prevention measures	Accident limitation measures
Technical measures	design of containment and service equipment according to normal conditions	accident-safe design with special equipment (e.g., double containment, fire protection, alarms, shut-down) special provisions (e.g., fail-safe, fire-safe, redundancy, diversity)	separation, size limitation, safety distance, hold-back measures
		marking, labeling	
Operational measures	experience, training		
	handling and testing instructions, maintenance		emergency response, recovery action
Quality assurance measures			

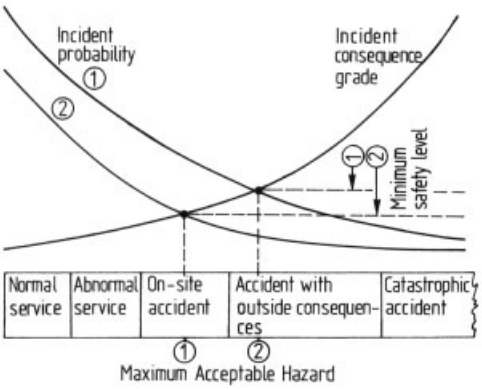

Figure 1. Correlation between incident probability, consequences, and safety level for two cases of acceptable hazard

4.1. Technical Safety Measures

Technical safety measures have to assure safe service conditions or (if they are so designed) safety in accidents. The design of a technical system has to consider the right configuration of "function units" with components made from suitable materials, correctly dimensioned, and equipped with all necessary devices. The design or construction process has to consider all relevant design loads:

Design load
Mechanical impacts
Pressure
- internal and/or external over-pressure
- design pressure
- working pressure
- test pressure
- hydrostatic pressure
- pump pressure

Vibration
Operational loads
Wind, snow floods
Accident impact
- drop
- crush
- puncture
- ripping
- missile

Thermal impact
Substance temperature
Ambient temperature
Fire
- torch fire
- full engulfment
- partial engulfment
- radiation from external fire

Corrosion
Substance
Environment
Time-induced effects
Aging
Creep
Stress relaxation

Construction/design features
Purpose
Lifetime
Materials
Dimensions
Equipment
Configuration of units and components
Safety principles
- safety level
- single or multiple barriers

redundancy
diversity
fail-safe design
fire-safe design

The following examples illustrate the interdependance of design features, design loads, and purpose.

4.1.1. Design and Equipment of Tanks and Packages

Transport and storage containers exceeding a capacity of 3 m^3 are usually designed as pressure vessels. Commonly, calculation of the resistance to internal over-pressure is based on the maximum vapor pressure of the dangerous substance under service conditions; this pressure depends on the kind of substance and the reference temperature of the vapor pressure. The vessel's resistance to internal over-pressure is determined by the wall thickness and the maximum allowable shell stress value, i.e., a specified proportion of the yield or tensile strength of the tank material. The ratio of strength to allowable stress is the safety factor: the material's ability to withstand mechanical impacts should only be consumed partially under normal service conditions, leaving a reserve in the case of accidents. The calculation of the minimum shell thickness for a cylindrical pressure vessel is given by:

$$s = \frac{D_a p}{20 \frac{K}{S} v + p} + c_1 + c_2$$

with s = minimum tank shell thickness (mm); p = design pressure (bar = 100 kPa); D_a = external diameter of the cylindrical tank shell (mm); K = material strength at design temperature (N/mm^2); S = safety coefficient at design pressure; v = "weld coefficient", representing the absorption of the allowable design stress in a welded joint (v = 1 if a 100 % nondestructive test of the welded joint is required); c_1 = allowable tolerances of the tank shell material's dimensional standards; and c_2 = possible effects of corrosion.

Other components of a pressure vessel (heads, manholes, flanges, pipes, valves, etc.) are designed according to similar calculations.

The whole vessel and piping design, including calculation, material selection, equipment, manufacturing, and testing must be carried out according to an accepted pressure vessel design code, e.g., ASME [17], or *AD-Merkblätter* [18].

The safety level expressed by these codes is adequate only up to the pressure hazard potential, usually expressed as the product of pressure and volume. For a given material, the wall thickness may be taken as indication for the safety level of a pressure vessel; every parameter determining this wall thickness (for a given design pressure) depends on the type of vessel material. These parameters have been established on the basis of experience for the most common structural materials, and are defined in the codes. For example, Table 3 shows safety coefficients for some ductile metals; interpretation of these data yields some valuable safety principles:

1) The better the ductility of the material, the lower the safety coefficient to be applied (plastification before fracture can give significant energy absorption before rupture of the vessel)
2) Vessels of forged or rolled steels and aluminum alloys, i.e., metals with reasonably high ductility, are stressed at the final inspection, and in reinspection tests, at a test pressure 1.3 times the maximum allowable working pressure (MAWP); the tank shell is stressed up to 90 % of the material's strength, usually the yield strength
3) Vessels of metals with limited ductility (e.g., cast iron) have to withstand a test pressure which is twice the MAWP

An additional possibility to influence the safety level via the tank shell thickness is the determination of the design pressure of the tank. Usually, the design pressure is the maximum over-pressure that may occur during operation. For liquids of low vapor pressure, the MAWP may be mainly determined by the filling pump pressure; for high vapor pressures (e.g., in transportation and storage of gases) the design pressure is the vapor pressure at the maximum operational temperature. These "reference temperatures", together with the relations between MAWP and test pressure, and the definitions of the maximum allowable stresses for different modes of transport and for storage are given in Figure 2.

Table 3. Safety coefficients for ductile metals

Material	Safety coefficient S at design temperature and maximum working pressure	Safety coefficient S' at test pressure
Forged and rolled steels	1.5	1.1
Cast steels	2.0	1.5
Ductile cast iron (DIN 1693)		
GGG 60/70	5.0	2.5
GGG 50	4.0	2.0
GGG 40	3.5	1.7
GGG 40.3 (with guaranteed toughness)	2.4	1.2
Aluminum and Al alloys	1.5	1.1

For the same tank size and material, calculation for a pressure-liquified gas such as propane shows that a road transportation tank with sun protection has to be designed with a shell 1.33 times thicker than that of an above-ground storage vessel. Transportation regulations generally have to consider a higher probability of accidents impacting the tank from the outside, e.g., in crashes. For hazardous liquids, this is achieved by setting minimum shell thickness standards or "fictive" high design pressures, to design a tank that is more resistant to external forces.

According to the specific hazard potential of substances, land transportation regulations (RID/ADR) require tank shell thicknesses based on "fictive" design over-pressure steps of 0.4, 1, 1.5, and 2.1 MPa; sea transportation regulations (IMGD-Code) require minimum shell thickness values of 6, 8, 10, or 12 mm (mild steel, tank diameter > 1.8 m). These regulatory shell thickness requirements qualitatively enhance tank safety, but only in a very general manner, and do not consider quantitatively safety against external impact. Attempts to increase the safety of tanks by technical measures and material selection are made in risk assessment studies [19, 20], or by investigations of the global or local resistance of tanks or tank components to accidental mechanical forces [21, 22].

The safety of packagings and intermediate bulk containers (IBC) has to be proved mainly by performance tests. According to the UN Test Requirements for Packagings (with the exceptions of Class 2 and Class 7 substances) the following tests have to be carried out on every design type:

Drop Test. The number of test samples and the drop orientations depend on the kind of packaging; the packaging must be dropped onto a rigid, nonresilient, flat, horizontal target. The drop height depends on the packing group (PG) of the substance, i.e; the specific hazard potential: 1.8 m for PG I, 1.2 m for PG II, and 0.8 m for PG III.

Leakproofness test at an air pressure not less than 30 kPa (PG I) or not less than 20 kPa (PG II, III).

Internal pressure (hydraulic) test at a test pressure of 250 kPa (PG I) or not less than 1.5 times the vapor pressure at 55 °C (minus 10 kPa), but at least 10 kPa (PG II, III).

Stacking test with minimum stacking height of 3 m.

A very special case is the set of performance test criteria for so- called type-B packages for radioactive materials. These criteria include tests of the ability to withstand severe accident conditions in transport. A package sample is subjected to the cumulative effects of mechanical tests (9 m drop onto an unyielding target, 1 m drop onto a steel bar, and a crush test by a 500 kg mass dropped from a height of 9 m) as well as a thermal test (30 min full fire engulfment, e.g., in a kerosene pool fire with flame temperatures exceeding 800 °C). An additional requirement is a water immersion test (150 kPa external gauge pressure of 2 MPa for packages with a radioactive inventory exceeding a specified limit). These severe performance tests, after which a specified activity release rate has to be guaranteed, lead to a package design that is safe, even under the worst accident conditions.

4.1.2. Fire Protection Measures

The biggest losses in industrial sites are often caused by fires. Storage sites or tanks with large amounts of hazardous materials have to be carefully fire protected. The usual classification is:

"Passive" fire protection measures, such as buildings, building components, insulation layers, and coatings resistant over a specified duration to a specified fire scenario, e.g., a furnace test fire [23]

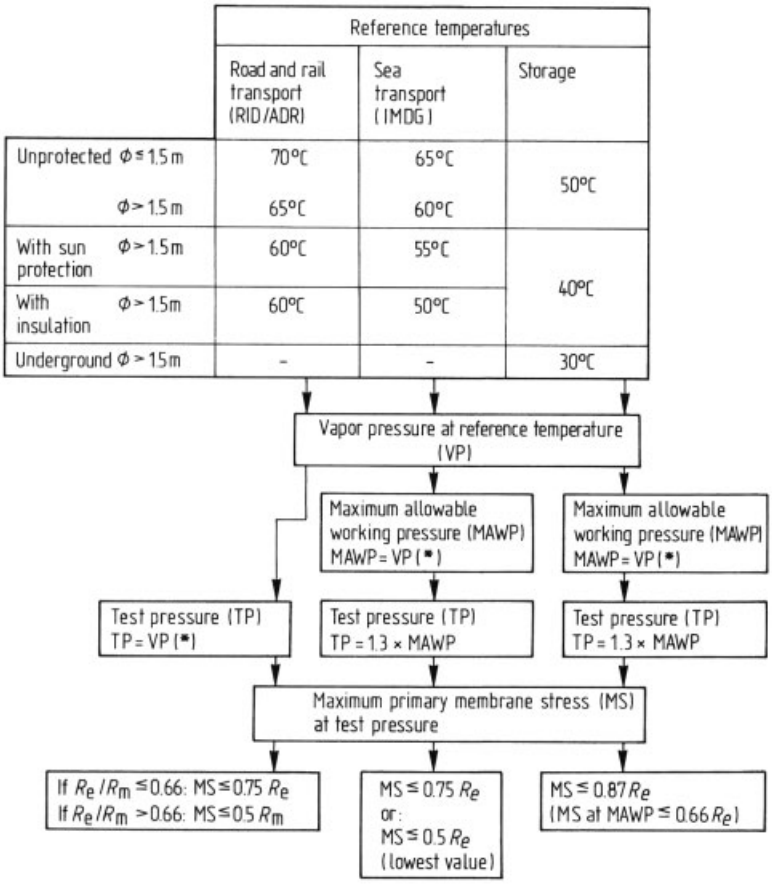

Figure 2. Pressure design criteria for gas tanks
(∗) Set pressure of the pressure-relief device.

"Active" fire protection measures, such as fire extinguishing systems, sprinkler and water spray systems

For effective fire protection of dangerous goods tanks, and for fire fighting in sites where the tanks are on fire or close to it understanding the behavior of tanks engulfed by a fire is essential. For LPG tanks Figure 3 shows the relation between fire duration and vessel-bursting pressure (upper curves), and the internal over pressure (lower curves). If the tank is equipped with a pressure-relief valve, and protected by an effective coating, which may be a layer of insulation, intumescent or subliming material [24], the internal over-pressure is always well below the bursting pressure. An unprotected tank, although equipped with a high- capacity pressure-relief device, will burst after a reasonably short fire duration if the tank is fully engulfed. If the tank is only partially engulfed this will happen just as fast if the fire engulfs the tank shell in the vapor space, where the shell is heated very rapidly; the shell wetted by the liquid is not heated so rapidly because of efficient heat transfer to the liquid. Figure 3 also shows cases where the tank has little or no pressure-relief apparatus; vessel bursting may occur very rapidly, in the case of LPGs leading to a boiling liquid expanding vapor explosion (BLEVE) with desastrous consequences for the environment (tank fragmentation, fire-ball heat radiation, and explosion overpressure).

Fire extinguishing systems, commonly used in closed-storage facilities, put the storage space under inert gas (e.g., CO_2) or foam to extinguish a fire. Sprinkler and water spray systems oper-

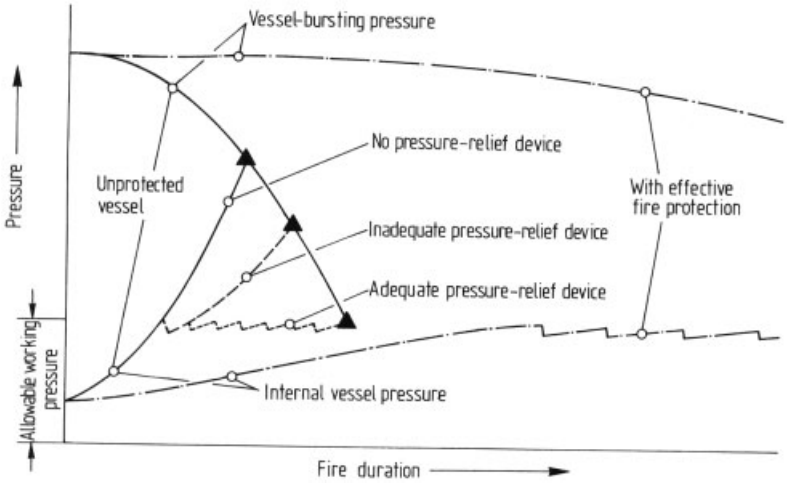

Figure 3. Liquefied petroleum gas (LPG) tank behavior in a fire
△ = BLEVE (boiling liquid expanding vapor explosion)

ate by cooling. All these systems have to be initiated, usually automatically by fire detectors; their reliability depends on the proper functioning of many components, and their full efficiency is reached only after a lag time. Therefore, inherently safe fire protection measures, e.g., a fire-resistant construction, is often preferred for sites with major risks. Details of fire protection methods can be found in industrial handbooks e.g., [25, 26].

4.2. Quality Assurance Measures

Quality assurance (QA) programs should be established for the design, manufacture, testing, documentation, use, maintanance, inspection, transport, handling, and storage operations ofall containments for dangerous goods to ensure compliance with the relevant safety regulations and the provisions of approval certificates. In this context, "quality" is the compliance of a tank or package at any stage of its lifetime with the regulations, so that an approved safe condition is achieved.

For a package, a brief survey of QA is shown in Figure 4.

Quality assurance is based on a quality policy, quality management (within all institutions that design, manufacture, operate, test, examine, inspect, or approve), and a system of quality controls and audits.

Basic ISO QA standards (with comparable European standards) are given in:

– ISO 9000 (EN 29 000) Quality Management and Quality Assurance Standards – Guidelines for Selection and Use
– ISO 9001 (EN 29 001) Qualitiy Systems – Model for Quality Assurance in Design/Development, Production, Installation and Servicing
– ISO 9002 (EN 29 002) Qualitiy Systems – Model for Quality Assurance in Production and Installation
– ISO 9003 (EN 29 003) Quality System – Model for Quality Assurance in Final Inspection and Test
– ISO 9004 (EN 29 004) Quality Management and Quality System Elements – Guidelines

Quality can be produced only by skilled engineering, manufacture, and operation, by well-educated and trained personnel. A very sophisticated QA program is able to filter out failures only to a certain extent.

4.3. Safety Distances

Accidental release should be prevented by technical and operational safety measures as described above, but within a safety analysis for

Phase	Item		Quality assurance tasks	Quality assurance instruments
Design control and approval	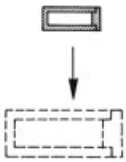	Test object	Quality control of: test object properties test parameters	Test program, test records, examination and documentation of test object properties, test report
		Package design	Validation of design safety evaluation, considering: transferability of the test results, assessment of quality standards	Safety evaluation report, including design specifications (drawings, specifications of materials and components)
Manufacture	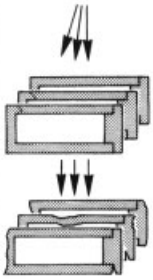	Package produced	Assessment of fabrication and test procedures Quality control to assure design specifications Evaluation of deviations from specification	Quality assurance program QA manual fabrication and inspection plans including: fabrication instructions test instructions responsibility of inspectors/experts according to a classification system final inspection plan QA records
Operation			Maintenance, safe operation, and quality control to assure design specifications	Operation and inspection plan Maintenance and reinspection test instructions } Transport Acceptance test criteria Storage manual } Storage

Figure 4. Quality assurance

any site where a major risk exists, the consequences of a potential hazardous material release have to be determined.

Estimation of consequences may be necessary for the planning of emergency response action, as well as for plant construction or storage size limitation to provide appropriate distances between the hazard source and other plant areas (storage spacing) or public facilities. First, a realistic accident scenario has to be considered to derive a source term (instantaneous release of a certain amount of dangerous goods, or a specified continuous release rate). Depending on the behavior of the substance in mind, the subsequent reaction may be ignition, explosion, release of hazardous reaction products, or dispersion only. Depending on the substance-specific hazard potential (specific toxicity, flammability, explosibility, causticity, radioactivity, etc.), and other influences (e.g., metereologic and plant conditions) the hazard spreads, up to a certain distance. The distance up to which the hazard is at an unacceptable level, e.g., 50 % lethal dose toxicity (LD_{50}), or heat radiation or explosion over-pressure causing severe injury or building damage is termed the "safety distance."

The effects of the release of flammable liquids and gases (heat flux, over-pressure and tank missile flight after a BLEVE, vapor cloud explosion, or fire) can be estimated from formulae in the appropriate guides and handbooks, e.g., [28, 30]. Somewhat more difficult is the calculation of gas dispersion to estimate, e.g., the lower flammability distance in LPG releases or toxic concentration distances (e.g., for ammonia or chlorine releases) because of the effects of metereologic circumstances; calculation standards for these are also available [31].

If the existing safety distance is less than an estimated unacceptable hazard distance, additional safety measures have to be taken. These may be either source-term-reducing measures, such as smaller vessels, thinner or shorter pipes between shut-off valves, or hazard-distance-reducing measures, such as gas-tight walls or confinements, or water curtains to reduce the gas concentration.

4.4. Emergency Response

Emergency response action is necessary as a back-up, to limit the consequences of an accidental release of hazardous substances in case of a failure of all technical barriers. The correct and durable marking, labeling, and placarding of dangerous goods containments is the first important measure for corrective action. Transport media for dangerous goods must carry additional labels or instructions identifying the primary and subsidiary risks of the substance, and indications of the initial emergency response. More details are necessary for the fire fighters at the accident site. Some countries have set up special task-forces with appropriate equipment for dangerous goods transport accidents; these task-forces can get additional support from quick-response hazardous substances data banks from specialist advisers, or from technical equipment supplied by the chemical industry. Support is organized either on a national basis, e.g., in Germany by TUIS, a transportation accident information and emergency response system run by the chemical industry [32], or internationally, e.g., in Europe by ICE run by the European Chemical Industry Council, CEFIC [33]. For major-risk storage sites, a detailed alarm and emergeny response plan that regulates all necessary action inside the plant, and the activation of external help by local authorities and fire fighters must be approved and regularly up-dated during operation.

5. Regulations and Standards

5.1. World Transport Regulations

The purpose of all regulations for the transport of dangerous goods is to minimize the associated risks, while limiting the costs for the industry concerned to an acceptable level.

The basis of all regulations is a classification system for dangerous goods by type of hazard and degree of danger.

According to the predominant hazard of the dangerous substance, classification is the same for all regulations for the transport of dangerous goods:

Class 1	Explosives
Class 2	Gases
Class 3	Flammable liquids
Class 4.1	Flammable solids
Class 4.2	Substances liable to spontaneous combustion
Class 4.3	Substances which in contact with water emit flammable gases
Class 5.1	Oxidizing substances
Class 5.2	Organic peroxides
Class 6.1	Toxic substances
Class 7	Radioactive material
Class 8	Corrosive substances
Class 9	Miscellaneous dangerous substances

About 2500 individual substances and articles have been classified; substances, not specifically listed in the regulations, but which are dangerous according to the criteria, have to be classified, to be acceptable for carriage under the same conditions as substances presenting a comparable hazard. Some substances are too dangerous for transport, and may be carried only in a stabilized form or in solutions or mixtures. All regulations specify the conditions of carriage for the dangerous substances, such as:

Conditions of packing
Marking and danger labels on packages
Information in the transport document
Conditions for tank transport
Mixed loading and segregation
Special conditions for different modes of transport

5.1.1. UN Recommendations on the Transport of Dangerous Goods

Since 1956 a Committee of Experts on the Transport of Dangerous Goods established by the United Nations Economic and Social Council has been developing Recommendations on the Transport of Dangerous Goods [34]: Although the UN Recommendations have no legal force, they aim at presenting a basic scheme to allow international regulations for the various modes of transport to develop in a uniform fashion.

All bodies responsible for the international regulations have committed themselves to harmonize their regulations as far as possible with the principles of the UN Recommendations.

The most important part of the Recommendations is the *List of Dangerous Goods Most Commonly Carried*. A four-digit number is assigned

in the list to each substance or article. Unfortunately, this UN number gives no information about the hazards of the substance or article.

Classification of dangerous substances not individually listed in the Recommendations is possible under a great number of "generic" or "not otherwise specified" (N.O.S.) entries. Such classifications should be based on the results of tests, and on experience [35].

The UN Recommendations also cover provisions for packing, multimodal tank transport, consignment procedures, marking, and labeling.

The Committee of Experts gives no recommendations for the transport of radioactive materials. In 1959, the UN Economic and Social Council asked the International Atomic Energy Agency (IAEA) to prepare Regulations for the Safe Transport of Radioactive Material [36]. These Regulations, which are referred to in the UN Recommendations, have been introduced in all international regulations for the transport of dangerous goods.

5.1.2. The IMDG-Code

The International Maritime Organization in London has prepared requirements for the safe transport of dangerous goods by sea. These requirements were originally based on the *International Convention for the Safety of Life at Sea* (SOLAS) of 1960. To prevent marine pollution by the transport of dangerous substances, the requirements also include provisions based on the *International Convention for the Prevention of Pollution from Ships* (MARPOL) of 1973.

The requirements are laid down in the *International Maritime Dangerous Goods Code* (IMDG-Code, [37]). The IMDG-Code is recommended to governments for adoption as the basis for national regulations in pursuance of their obligations under the SOLAS Convention.

The IMDG-Code follows very closely the UN Recommendations with regard to the classification, packing, and tank transport of dangerous goods.

A special feature of the IMDG-Code is the presentation of provisions for packing: an individual page for each substance or entry makes the IMDG-Code the most voluminous of all international regulations for the transport of dangerous goods.

In addition to the provisions in accordance with the UN Recommendations, the IMDG-Code sets out special requirements for the transport of dangerous goods in ships, such as:

Provisions for freight container traffic
Stowage and segregation of dangerous goods on ships
Fire precautions
Dangerous goods on roll-on, roll-off ships
Transport of solid bulk materials.

Furthermore, a supplement to the IMDG-Code includes publications of the International Maritime Organization relating to the IMDG-Code, such as:

The Emergency Procedures for Ships carrying Dangerous Goods (EmS)
The Medical First Aid Guide (MFAG)
The Code of Safe Practice for Solid Bulk Cargoes
Reporting procedures
The Guidelines for Packing Cargo Transport Units
Recommendations on the use of pesticides in ships.

5.1.3. The ICAO Technical Instructions

The International Civil Aviation Organization in Montreal is responsible for regulating the air transport of dangerous goods. In 1984, the first edition of the ICAO *Technical Instructions for the Safe Transport of Dangerous Goods by Air* [38], based on Annex 18 to the Convention on International Civil Aviation was introduced. Before then, the transport of dangerous goods by air had been regulated by the International Air Transport Association (IATA), also in Montreal in the IATA Restricted Articles Regulations.

The ICAO Technical Instructions are also fully harmonized with the UN Recommendations as regards the classification and the provisions for packing, labeling, and documentation.

There are, however, many restrictions on air transport: many substances and articles from the UN Recommendations, e.g., explosives, are forbidden for air transport, and there are many restrictions on the net quantity per package. Restrictions are more severe for passenger than for cargo aircraft.

Most of the conditions of carriage in the ICAO Technical Instructions are included in the alphabetical Table of Dangerous Goods, which also refers to packing instructions and special provisions.

In addition to the provisions in accordance with the UN Recommendations, the ICAO Technical Instructions cover special requirements for air transport, such as:

Provisions for magnetized material
Loading of dangerous goods on aircraft
Information to flight crews
Training of shippers and operators and all other persons handling dangerous air cargo
Variations of the Technical Instructions, notified by states and airline operators.

5.2. Regional Transport Regulations

5.2.1. Europe

The first regulations for international transport of dangerous goods in Europe were developed over 100 years ago for rail transport. Regulations for road transport and transport by inland waterway were developed after World War II, and are all based on the regulations for rail transport (RID). Considerable effort has been made to harmonize all European regulations on aspects such as classification, packing, labeling, and marking as well as the requirements for tank wagons, tank containers, and tank vehicles.

All European regulations are being harmonized with the UN Recommendations. This is difficult, because there are still fundamental differences in safety principles underlying the UN Recommendations and the European regulations.

The Regulations Concerning the International Carriage of Dangerous Goods by Rail (RID). The RID (*Règlement Concernant le Transport International Ferroviaire des Marchandises Dangereuses* [39]) is Annex I to the International Convention concerning the Carriage of Goods by Rail (CIM). The RID is applicable to transport between most European countries and some countries of northern Africa and the Middle East, i.e., the Member States of the Intergovernmental Organization for the International Carriage by Rail (OTIF), based in Bern.

The first regulations for the transport of dangerous goods by rail came into force in 1893. Development of the regulations is the responsibility of the RID Safety Committee, which works closely with road transport experts in the joint RID/ADR Meeting.

Although the criteria for classification of dangerous goods in RID are harmonized with the UN Recommendations, there is a special feature in RID and other European regulations. The conditions for carriage are set out for the substances of each class separately, starting with a systematical list of substances and articles; all substances and articles are grouped under collective headings according to their physical and chemical properties and their degrees of hazard. The appropriate N.O.S. entry from the UN Recommendations concludes every group of substances under the collective heading, to accommodate classification of new substances and articles. This classification system ensures that the same conditions of carriage apply to substances under one collective heading or item; the format of the regulations can be kept simple, without the need for amendment of the regulations for every new substance.

The conditions of carriage in each class cover the following subjects:

– Conditions of packing, including marking and labeling
– Particulars in the consignment note
– Conditions for wagons and loading, including marking and labeling

The RID also includes appendices with provisions for testing substances, packaging (including intermediate bulk containers), for radioactive material, and provisions for tank containers and tank wagons.

The European Agreement Concerning the International Carriage of Dangerous Goods by Road (ADR). The ADR (*Accord Européen Relatif au Transport International des Marchandises Dangereuses par Route* [40]) was concluded in 1957. Most European states are contracting parties to the ADR. Development of the ADR is the responsibility of a Working Party, established by the Inland Transport Committee of

the United Nations Economic and Social Council.

Annexes A and B cover the same subjects as RID. Annex B also covers some provisions important for road transport, such as:

Technical equipment of road vehicles

General service requirements, driver training

Provisions concerning loading, unloading, and handling

Provisions concerning the operation of vehicles

The European Provisions Concerning the International Carriage of Dangerous Goods by Inland Waterway (ADN). It is the intention of the UN Economic Commission for Europe to additionally put into force an agreement for the transport of dangerous goods by inland waterways. At present the ADN exists only in the form of a draft.

The Regulations for the Carriage of Dangerous Goods on the Rhine (ADNR). The ADNR (ADN-Rhine) is based on the existing drafts for the ADN and has been developed by the Central Commission for the Navigation on the Rhine in Strasbourg. The plan is for the ADNR to be subsumed by the ADN.

For the transport of packaged goods, the ADNR refers to other regulations for the transport of dangerous goods; it mainly regulates the transport of dangerous goods in bulk, and therefore consists of provisions for the construction, equipment, and operation of vessels.

Directives of the European Communities with Regard to the Transport of Dangerous Goods. It is the intention of the Commission of the European Communities to extend the implementation of the existing international regulations for the transport of dangerous goods to national transport in the Member States of the Community.

5.2.2. United States

The transport of dangerous goods (called hazardous materials in the United States) is subject to Title 49 of the United States Code of Federal Regulations Parts 107 – 180 (49 CFR Parts 107 – 180), except that bulk transport in tank ships and tank barges on the navigable waters of the United States is covered by other regulations. The Associate Administrator for Hazardous Materials Safety of the Research and Special Programs Administration (RSPA) is responsible for maintaining the regulations and is also the United States competent authority for purposes of the ICAO Technical Instructions on the Safe Transport of Dangerous Goods by Air and the IMDG-Code.

In many respects the contents of the United States Hazardous Materials Regulations (HMR) are similar to the international requirements for dangerous goods transport contained in the ICAO TI and the IMDG-Code. This is largely the result of amendments made to 49 CFR in December 1990 and 1994. These amendments substantially aligned United States requirements for classification, transport documents, nonbulk packaging, labeling, package marking, and placarding with the UN Recommendations on the Transport of Dangerous Goods. In addition, the HMR also include vehicle construction and operating requirements for transporting hazardous materials by rail and highway, operating requirements for transporting hazardous materials by air, and the procedures under which an exemption to the regulations may be obtained.

In the case of consignments by air or water, two provisions in the Regulations are particularly important. Subject to specified conditions, 49 CFR 171.11 authorizes transport of hazardous materials on aircraft and any associated highway transport in accordance with the ICAO Technical Instructions. Similarly, 49 CFR 171.12 authorizes transport of hazardous materials by ship and any associated land transport in accordance with the IMDG-Code if part of the transport of the material involves transport by water. A number of exceptions apply to transport in accordance with the ICAO and IMO requirements and the reader is advised to consider these provisions which are identified in 49 CFR 171.11 and 171.12. Some of the more important conditions include:

Environmentally hazardous substances are identified in the Regulations and must be transported under UN 3077 or UN 3082 if they do not meet the criteria of hazard Classes 1 – 8, and special hazard communication requirements apply to environmentally hazardous substances identified in the regulations.

Transport of hazardous materials in portable tanks is subject to 49 CFR which contains its own tank construction requirements.

Materials that meet the criteria of Division 6.1 on the basis of their toxicity by inhalation are referred to as "poison by inhalation" (PIH) materials which are subject to special packaging, marking, labeling, and transport document requirements. It should be noted that a number of substances have been found to meet the criteria of PIH materials, even though this is not reflected by their UN classification.

Gas cylinders used in transport must meet construction requirements in 49 CFR.

Explosives, including materials that exhibit explosive properties, require special authorization by RSPA.

Emergency response information must accompany hazardous materials shipments, and the transport document must include a telephone number where additional information on the material may be obtained at any time while the material is in transport.

Training requirements apply to shippers and carriers that bring hazardous materials into the United States

Additional restrictions pertaining to the transport of radioactive materials are included.

The requirements also include special provisions relating to air transport.

Under 49 CFR 171.12 (c), fewer special provisions apply when hazardous materials transported in accordance with the IMDG-Code are transported within a single United States port area.

5.2.3. Japan

Japanese regulations for transport of dangerous goods by rail and road exhibit some differences from those of European countries and the United States. International carriage of dangerous goods by sea and/or air is subject to the domestic regulations based on the IMDG-Code and ICAO Technical Instructions, but for inland transport by rail and/or road, the current regulations are not based on United Nations or other international recommendations. Dangerous goods can be transported from foreign countries to Japan by ship or aircraft under common classification, packaging, labeling, marking, etc., in compliance with IMDG-Code or ICAO Technical Instructions, but for inland transport by rail and/or road, carriage is regulated by Japanese requirements. In Japan, dangerous goods are regulated en bloc for manufacturing, storage, transportation, consumption, etc., except for regulation of carriage by ship and aircraft.

Japanese regulations for transport of dangerous goods are summarized as follows.

Sea transport is regulated by the following laws:

1) Ships Safety Law. Transport requirements for dangerous goods by ship are provided in the Regulations for Carriage and Storage of Dangerous Goods by Ship as an order of the Ships Safety Law, in which almost all of the IMDG-Code has been adopted.
2) Port Regulation Law. This provides for berthing and cargo handling of a ship carrying dangerous goods in port, but not packaging, labeling, marking, etc., which must be in accordance with the Ships Safety Law.

Air transport is regulated by the Civil Aeronautics Law. This prescribes transport requirements regarding classification, packaging, labeling, marking, etc., and complies the requirements of ICAO Technical Instructions.

Inland transport of dangerous goods is implemented in Japan under comprehensive regulations regarding manufacturing, storage, transport, consumption, etc. Classification, packaging, etc. for transport are not based on international recommendations.

1) Fire Service Law. This provides requirements for storage, road tank vehicles, means of road transport, emergency procedures, etc. It covers Classes 3, 4 and 5 of the UN Recommendations, with some differences in classification criteria.
2) Railway Operation Law. Requirements for carriage of dangerous goods by rail are regulated by this law, but classification, packaging, labeling, marking, etc. are subject to other inland transport regulations, such as the High-Pressure Gas Control Law and the Explosives Control Law. It covers all classes of the UN Recommendations, with some differences in classification criteria.

3) High-Pressure Gas Control Law. Storage, packaging, transport, etc. of gases on land are regulated by this law. Classification criteria are different from the UN Recommendations.
4) Explosives Control Law. Manufacturing, storage, transport, sale, consumption, etc. of explosives are regulated by this law. The Classes and Divisions of the UN Recommendations are not adopted.
5) Poisonous and Deleterious Substances Law. Import, manufacturing, sale, handling, transport, etc. of poisonous and irritant substances are regulated by this law. It covers Classes 6.1 and 8 of the UN Recommendations, with some differences in classification criteria.
6) Law for the Regulation of Nuclear Source Materials, Nuclear Fuel Materials, and Reactors. This provides land transport requirements, e.g., packages, test procedures, radiation protection, for nuclear materials. It adopts IAEA Regulations (SS 6).
7) Law for the Prevention of Damage from Radioactive Isotopes. This law provides land transport requirements, e.g., packages, test procedures, radiation protection, for radioactive materials other than nuclear materials. It adopts IAEA Regulations (SS 6).

6. Containment Technology

6.1. Transport of Piece Goods

Shipment of dangerous goods in packed form, i.e., in packagings or intermediate bulk containers (IBC), is still the most common means of carriage from one site to another by all modes of transport.

6.1.1. UN Packagings, "Packagings by Description", Intermediate Bulk Containers

In its *Recommendations on the Transport of Dangerous Goods* [41], the UN Committee of Experts on the Transport of Dangerous Goods has issued a set of provisions to provide as much safety for the transport of dangerous goods as possible with regard to the needs of all parties concerned, e.g., administrators, commercial companies, or the public.

The recommendations on the packing of dangerous goods form the basis for existing international [42–45] and national regulations. Account is also taken of a prevailing trend to replace the detailed specification/description by tests designed to ensure that the packaging or IBC can withstand normal conditions of transport, and to ensure a desirable level of safety.

6.1.1.1. UN Packagings

Types of Packagings. For the so- called UN Packagings, the types of packaging are defined as follows:

Bags: flexible packagings made of paper, plastic film, textiles, woven material, or other suitable materials.

Boxes: packagings with complete rectangular or polygonal faces, made of metal, wood, plywood, reconstituted wood, fiberboards, plastics, or other suitable material.

Drums: flat-ended or convex-ended cylindrical packagings made of metal, fiberboard, plastics, plywood, or other suitable material. Wooden barrels of jerricans are not covered by this definition.

Jerricans: metal or plastics packagings of rectangular or polygonal cross-section.

Wooden barrels: packagings made of natural wood, of round cross-section, having convex walls, consisting of staves and heads, and fitted with hoops.

There are two more complex forms of dangerous goods packagings:

Combination packagings: consist of one or more inner packagings secured in an outer packaging in such a way that they cannot break, be punctured, or leak their contents into outer packagings (usually boxes).

Composite packagings: consist of an outer packaging and an inner receptacle, constructed so as to form an integral packaging. Once assembled, it remains thereafter an inseparable unit; it is filled, stored, transported, and emptied as such.

General packing requirements applicable to the packing of substances of all classes other than Classes 2 and 7 define the design and behavior of packagings. It is required that these

should be constructed and closed so as to prevent any leakage which might occur under normal conditions of transport, by vibration, or by changes in temperature, humidity, or pressure. Parts of packagings which are in direct contact with dangerous substances should not be affected by chemical or other action of those substances, or should be provided with a suitable inner coating or treatment, e.g., a liner.

The packagings should be manufactured and tested under a quality assurance program which satisfies the competent authority that each manufactured packaging meets the requirements of the design type.

Testing and Approval. Requirements in the regulations for package strength are expressed as performance standards or tests. The regulations prescribe what must be achieved, rather than how itis done. Therefore, the most important item is that each packaging, except inner packagings of combination packagings, should conform to design types successfully tested, as set out in the regulations and in accordance with procedures established by the competent authority. The testings consist of drop tests (for all design types) and stackings tests (for all design types except bags and nonstackable composite packagings). Packagings intended to carry liquids, i.e., drums, jerricans, additionally have to pass hydraulic pressure tests and leakproofness tests; if they are made of plastics they also have to prove the compatibility of the plastics materials with the contents, i.e., prestorage for six months with the liquid(s) intended or with so-called standard liquids: wetting solution, acetic acid, *n*-butyl acetate/*n*-butyl acetate-saturated wetting solution, mixtures of hydrocarbons (white spirit, nitric acid, or even water). The latter standard liquids only apply to defined high molecular mass polyethlene packagings. These are the so-called compatibility tests.

Tests must be carried out on packagings and packages prepared as for despatch, including inner packagings of combination packagings. The packagings shall be filled to not less than 95 % of their capacity for solids or 98 % for liquids. A given number of test samples have to pass the tests in such way that the weakest part of the packaging is tested. The severity of the test conditions depends on the packing group(s) (PG) for which the packaging is intended.

There are drop tests from three different heights: 1.8 m for PG I, 1.2 m for PG II, or 0.8 m for PG III. Where the relative density of a liquid to be carried is greater than 1.2 g/cm^3, the drop height (in meters) is calculated on the basis of the relative density by multiplying it by 1.5 for PG I, by 1.0 for PG II, and by 0.67 for PG III. The drop tests are carried out on a target with a rigid, nonresilient, flat, and horizontal surface, having mass 50 times that of the test specimen (see ISO 2248).

There are two different air pressures for the leakproofness test, i.e., not less than 30 kPa for PG I or not less than 20 kPa for PG II and III.

The hydraulic (or internal) pressure test to be applied depends on the vapor pressure of the filling substance, normally at 55 °C, but must not be less than 100 kPa.

The stacking test must prove that the test samples can withstand an additional mass equivalent to a stacking height of at least 3 m.

There are additional tests such as the permeability test for plastic drums and jerricans when used for flammable liquids with flash point < 55 °C, or a supplementary test for wooden barrels.

The competent authority may also establish procedures for the approval of packagings, now common practice in countries which have adopted the UN Recommendations or international regulations derived therefrom into national law.

Coding System and Marking. In order to identify the packagings with their design and tested properties, a system of coding and marking has been developed. The code consists of an arabic numeral indicating the kind of packaging, e.g., 1 for a drum, 3 for a jerrican, 4 for a box, or 5 for a bag, followed by one or more capital letters indicating the nature of the material, e.g., A for steel, G for fiberboard, H for plastics, etc., followed where necessary by an arabic numeral indicating the category of packaging, e.g., 1 when the packaging is designed with a nonremovable head or a closure with diameter ≤ 70 mm, 2 for a removable head or a closure with diameter > 70 mm. There are additional coding provisions for composite packagings and packagings manufactured to different specifications.

The complete marking of a dangerous goods packaging, having met the UN Recommendations and been approved by a competent authority, normally consists of the UN symbol followed by the packing code; a code in two parts, i.e., the letter X, Y, or Z designating the packing groups I, II, or III for which the design type has been approved and the relative density (liquids) or the maximum gross mass in kilograms; either a letter S (solids or inner packagings) or, where a hydraulic pressure test has been passed, the test pressure in kPa; the year of manufacture (last two digits); the mark of the state in which the approval was issued; and an identification mark specified by the competent authority.

Typical markings are for example:

UN 1A1/X/250/94/D/BAM 1234-00, where 1A1 means a steel drum with openings of diameter < 70 mm, submitted to drop tests from a height of 1.8 m, and tested at an internal hydraulic pressure of 250 kPa. The certificate of approval has been issued for PG I by the *Bundesanstalt für Materialforschung und -prüfung* (BAM) in Berlin, being the competent authority in Germany, the registration number is 1234, and the drum was manufactured in 1994 by a company with registration letters 00. This steel drum can be used for liquids of packing groups PG I, II, and III.

Or another example:

UN 4G/Y 20/S/94/D/BAM 9876-QQ, where 4G means a fiberboard box, submitted to drop tests from 1.2 m. The box can be used for solids of PG II and III or inner packaging(s) with substances of PG II and III. Here too the certificate of approval has been issued by the BAM in Germany; the registration number is 9876, and the box was manufactured in 1994 by the manufacturer with the registration letters QQ.

6.1.1.2. Packagings by Description

Though the system of tested and approved design types of packagings has been established for dangerous goods of Classes 1, 3, 4.1, 4.2, 4.3, 5.1, 5.2, 6.1, 6.2, 8, and 9, there are still so-called described packagings, especially for Class 2, and in some cases 7. There are provisions for materials properties, e.g., aluminum alloy receptacles for certain gases of Class 2, or steel receptacles. The same applies for packagings intended for the shipment of low-specific-activity radioactive material (LSA-I) of Class 7, where provisions for general conditions of packagings also apply with no particular requirements for testing.

Class 2 is under revision, and it is planned to establish provisions for the testing of the appropriate packagings.

6.1.1.3. Intermediate Bulk Containers (IBC)

An intermediate bulk container (IBC) is a rigid or flexible portable packaging, other than specified in Section 6.1.1.1, that has capacity ≤ 3000 L (≤ 250 L for sea transport) is designed for mechanical handling, and is resistant to the stresses produced in handling and transport as determined by specified tests. Typically, IBCs are intended for multimodal transport of PG II and III substances.

Types of IBC. *Metal IBCs* consist of a metal body together with appropriate service and structural equipment.

Flexible IBCs consist of a body constituted of film, woven fabric, or any other flexible material, or combination thereof, together with any appropriate service equipment and handling devices.

Rigid plastics IBCs consist of a rigid plastics body, which may have structural equipment together with appropriate service devices.

Composite IBCs with plastics inner receptacles consist of structural equipment in the form of a rigid outer casing enclosing a plastics inner receptacle, together with any service or other structural equipment. They are so constructed that the inner receptacle and outer casing once assembled form, and are used as such, an integrated single unit to be filled, stored, transported, or emptied as such.

Fiberboard IBCs consist of a fiberboard body with or without separate top and bottom caps, if necessary an inner liner (but no inner packagings), and appropriate service and structural equipment.

Wooden IBCs consist of a rigid or collapsible wooden body, together with an inner liner (but no inner packagings) and appropriate service and structural equipment.

The IBCs must be so constructed and closed that none of the contents can escape under normal conditions of transport. All service equipment must be so positioned or protected as to minimize the risk of leakage owing to possible damage during handling and transport. The IBCs must be designed, manufactured, and tested under a quality assurance program which satisfies the competent authority, in order to ensure that each manufactured IBC meets the given requirements.

Testing and Approval. The design type of each IBC must be tested and approved by the competent authority. For each design type, a single IBC must pass the required tests before such an IBC is used. An IBC design type is defined by the design, size, material, and wall thickness, manner of construction, and means of filling and discharging; it may include various surface treatments.

Test shall be carried out on IBCs prepared for despatch, i.e., filled either with the contents intended to be carried or with other substances having at least equivalent physical properties, e.g., mass, grain size, relative density. For the latter, the same calculation procedure used for packagings can be applied.

The main test provisions are:

– A bottom lift test for each design type to be handled in this way, except for flexible IBCs
– A top lift test for metal, flexible, rigid plastics or composite IBCs
– A tear test for flexible IBCs
– A stacking test for all design types
– A leakproofness test for IBCs intended to carry liquids
– An internal hydraulic pressure test for IBCs intended to carry liquids
– A drop test for all design types; another IBC of the same design may be used
– A topple test for flexible IBCs only
– A righting test when a flexible IBC is designed to be lifted from the top or from the side

Coding System and Marking. The code system for IBCs consists of two arabic numerals indicating the type of IBC, a capital letter indicating the nature of the material and, where necessary, an arabic numeral indicating the category of IBC within the type. For composite IBCs, two capital letters are used; the first indicates the material of the inner receptacle, and the second that of the outer packaging.

Rigid IBCs for solids are coded 11 if loaded and/or unloaded by gravity, 21 if loaded and/or unloaded under pressure > 10 kPa, and 31 if intended to carry liquids. Flexible IBCs are coded 13.

The nature of the material is indicated by A for steel (all types and surface treatments), B for aluminum, C for natural wood, D for plywood, F for reconstituted wood, G for fiberboard, H for plastics, L for textiles, M for paper (multiwall), and N for metals other than steel or aluminum; i.e., the letters are the same as for the materials of packagings.

The IBC code is followed in the marking by a letter indicating the packing group(s), i.e., Y for PGs II and III, Z for PG III. The use of IBCs for PG I substances is still under discussion, and up to now scheduled only for certain solids.

Deviating from the marking system of packagings, there is a so-called primary marking consisting of: the UN symbol; the design type code; the letter for the packing group(s); the month and year (last two digits) of manufacture; the mark of the state in which the approval was issued; the name or symbol of the manufacturer, or any other identification as specified by the competent authority; the stacking test load in kilograms; and the maximum permissible gross mass or, for flexible IBCs, maximum permissible load in kilograms.

Additional marking follows the primary one, namely the tare mass (kg) for all IBCs other than flexible IBCs; for metal IBCs, rigid plastics IBCs, and composite IBCs the capacity (L at 20 °C); the date of the last leakproofness test (month and year), if applicable; date of last inspection (month and year); and the maximum filling/discharge pressure (kPa), if applicable. For metal IBCs only, the marking is completed by indicating the body material and its minimum wall thickness (mm) and the serial number of the manufacturer. For rigid plastics IBCs and composite IBCs, the test pressure (gauge) in (kPa), if applicable, concludes the marking.

A typical marking of a metal IBC could be: UN 11A/Y/0394/NL/Van Hoog 001/5500/1500, where 11A means a steel IBC for solids discharged, e.g., by gravity, for PG II and III;

manufactured in March 1994; approved by The Netherlands; manufactured by van Hoog in conformity with a design type to which the competent authority has allocated the serial number 001; load used for the stacking test 5500 kg; maximum permissable gross mass 1500 kg.

Operational Requirements. Before being filled and handed over for transport, every IBC must be inspected to ensure that it is free from corrosion, contamination, or other damage, and with regard to proper functioning of any service equipment.

When filled with liquids, sufficient ullage shall be left to ensure that, at the mean bulk temperature of 50 °C, the IBC is not filled to more than 98 % of its water capacity.

During carriage, no dangerous residue should adhere to the outside of an IBC. Also, IBCs must be securely fastened or contained within the transport unit so as to prevent lateral or longitudinal movement or impact, and to provide adequate external support.

After discharging, an empty IBC must be treated as for a filled IBC, unless it has been purged of the residue of the dangerous substance.

Where IBCs are used for the transport of liquids with a flash point ≤ 60.5 °C (closed cup, see also page **??**) or powders liable to dust explosion, measures must be taken to prevent a dangerous electrostatic discharge.

6.1.2. Packagings for Radioactive Material

Required packaging for radioactive material (RAM) is based on the assurance that the level of safety built into such a packaging meets the potential hazards of the contents being shipped. Because of the specific dangers arising from RAM, there are special regulations to ensure their safe transport [46, 47].

There are four primary types of package, relating to the activity and physical form of their radioactive contents: excepted packagings, industrial packagings, type A packagings, and type B packagings.

An *excepted package* contains quantities of RAM sufficiently small (e.g., smoke detectors) to allow their exception from most design and use requirements. Such packages must, however, meet certain requirements which ensure that their contents can be identified when opened, and that they can be safely handled and transported. These include design requirements to assure safe handling and stowage of the package, and to exclude adverse effects of shock, vibration, collection and/or retention of water, and chemical or radiolytic degradation of the packaging materials.

Because of the extremely low hazard posed by the contents of excepted packages, they may be transported with limited administrative controls, and without externally visible warning labels. Local postal regulations may permit certain packages of this type to be shipped by mail.

Industrial packages are used to transport materials known as either low specific activity (LSA) or surface- contaminated objects (SCO). Materials of low specific activity, having little activity per unit mass, and certain nonradioactive objects, having low levels of surface contamination, are safe for two reasons: either the contained activity is very low, or the material is in a form which is not easily dispersed and which presents only a low internal radiation hazard.

Materials of low specific activity, such as radioactive ores, can sometimes be transported unpackaged. Otherwise, these materials are transported in industrial packages.

In addition to meeting the requirements for excepted packages, industrial packages must meet requirements relating to normal conditions of transport which include minor mishaps. Industrial package integrity is graded in relation to the hazard posed by its contents. There are three types:

Industrial packages type 1 (IP-1) must meet certain temperature and pressure requirements beyond those imposed on excepted packages
Industrial packages type 2 (IP-2) also have to pass free drop and stacking tests
Industrial packages type 3 (IP-3) must in addition meet water spray and penetration test requirements

Type A packages are intended to provide a safe, economical means to transport relatively small quantities of radioactive material: they are expected to retain their integrity under "normal" abuse, i.e., quite likely to occur during transport:

Falling from vehicles or being dropped from similar heights
Being struck by a sharp object which may penetrate the surface
Being exposed to rain
Having other cargo stacked on them

These packages must be designed to satisfy all the requirements imposed on an IP-3 package, and must also meet more stringent test requirements if the radioactive content is liquid or gaseous. They must also satisfy stringent dimensional, ambient environment, internal pressure, and containment specifications which are not imposed on industrial packages.

It is assumed that Type A packages may be damaged in a severe accident, and that a fraction of their contents may be released. The regulations therefore prescribe limits on the maximum amounts of radionuclides that can be transported in such packages. These limits ensure that in the event of a release the risks from external radiation or contamination are low.

In addition to normal environmental conditions, such as high and low pressure, and to reduced atmospheric pressure as sustained by the excepted and industrial packages, Type A packages must be able to retain their contents and shielding integrity if subjected to:

1) The free drop test, where the test specimen is dropped onto a nonresillient hard target so as to cause maximum damage to the safety features to be tested. The normal drop height is 1.2 m. Heavier packages, of more than 1.5 t are dropped from a lower height ($\geq$0.3 m). Where a package is intended to carry liquids or gaseous contents the drop height is 9 m.
2) The compression (stacking) test, where the package is subjected for a period of 24 h to a load equal to or greater than five times the mass of the actual package.
3) The penetration test, where the packaging is placed on a rigid, flat, horizontal surface, and a bar of mass 6 kg and 3.2 cm diameter with a hemispherical end is dropped from a height of 1.0 m (1.7 m if the package is intended to carry a liquid or gas), with its longitudinal axis vertical, so that it falls onto the center of the weakest point of the package.

Type B packages are used to carry larger amounts of radioactive material. They must be able to withstand the effects of servere accidents; tests for resistance to immersion, representing hypothetical accident conditions, are required. Additionally, each design must be approved by the competent authority of the country in which the package was designed; and under some conditions, by the competent authority of each country through or into which it is shipped.

Type B packages are used for carrying radioisotopes for industrial radiography, irradiated nuclear fuel, nuclear wastes, and similar highly radioactive material. Type B packages must be capable of withstanding both Type A and Type B tests.

Packages Designed for Fissile Materials. This type of package is required if the contents are capable of sustaining a nuclear chain reaction. Special assessments and controls are required for such packages. And each design must be approved by the competent authority of each country from, through, or into which it is shipped.

Type A and B packages and packages for fissile materials also have to meet specific requirements. These are:

Mechanical tests, where any given packaging must pass two tests. In most cases, an impact (or drop) and a penetration (or puncture) test are required. These consist of dropping a test specimen onto two different targets so to cause maximum damage. In one drop test, the packaging is dropped from a height of 9 m onto a unyielding surface, in the other it falls onto a mild steel bar, 15 cm in diameter, from a height of 1 m. The bar is rigidly mounted, perpendicular to the target surface, and must project at least 20 cm above the base. An alternative to the 9 m drop test for some lightweight Type B packages is the so- called *crush-test*. This consists of placing the packaging on the unyielding target in such a way as to experience the maximum possible damage when a 500 kg steel mass is dropped onto it from a height of 9 m.

Thermal test, where the packaging is fully engulfed in a thermal environment of at least 800 °C for 30 min. Any combustion of the packaging's components must be allowed to continue self-extinction.

Water immersion test, where most packagings are immersed in water at a pressure equivalent to a depth of at least 15 m for not less than

8 h. A water immersion test applicable to some packagings for irradiated nuclear fuel requires immersion in water at a pressure equivalent to a depth of at least 200 m for not less than 1 h resulting in no leakage.

Means for identifying shipments of radioactive material are: labels (with the trefoil symbol); markings assigned by the competent authority – both labels and markings must be visible on the packaging; and/or placards (with the trefoil symbol) on vehicles, rail cars, freight containers, and portable tanks.

As outlined earlier, safety in transport depends primarily on the integrity of the packaging. In particular, packages designed for dangerous goods and especially those for the transport of materials with high potential hazards are designed and manufactured to the highest engineering standards to meet the requirements of the transport regulations. The regulations for the transport of radioactive materials were the first to contain specific provisions on quality assurance (QA) for all transport operations; QA is defined as any systematic evaluation and documentation of performance judged against regulatory requirements. Where approval of the competent authority is required for a design or shipment, such must be contingent on the adequacy of a QA program.

6.1.3. Consignments, Means of Loading, and Loading Security

The shipment of dangerous goods from one site to another on public routes involves a number of measures. These are covered by the so- called consignment procedures. The whole cycle from the manufacturing of the substance or product to its final use is a complex interaction of all parties and measures involved.

The manufacturer has to inform the consignor of any dangerous properties. The packer has to choose the admissible packaging or IBC, to take care of any prohibition of mixed loading, and to be responsible for the correct marking and labeling of the packages or IBCs. The consignor has to ensure proper documentation for the shipment, including the transport document containing at least: a description of the goods, including the substance identification number (where available); the class; necessary initials (if any); the number and a description of the packages or IBCs; the gross mass; the names and addresses of the consignor and consignee(s); and/or declaration required by the terms of any special arrangement. The carrier has to take care that all relevant and necessary information about the load and the transport documents be handed to the driver, that no damaged packages have been loaded, that the necessary testing certificates for tanks or tank containers are available, that filling limits have been observed, and all necessary labeling of the vehicle or container has been applied. There are additional requirements for the driver and assistant (if any) as well as for the vehicle itself [44].

The various components of a dangerous goods load must be properly stowed on the vehicle and adequately secured to prevent them from being displaced in any way in relation to each other and to the walls of the vehicle. Minor displacements may be permitted, provided they are of a stabilizing nature, and do not lead to constant frequent motions of the load in the cargo space. If the load comprises goods from different categories, the packages of dangerous goods have to be separated from the others. The load has to be uniformly distributed over the loading space, and/or sufficiently secured against slipping backwards, sideways, or falling out. Packages must be stowed so they are protected from any penetration or impact, and from being squeezed, punctured, or exposed to a pressure which might break the package. This can be achieved by clamping beams, brackets, pads, or dunnage baulkings. Cargo spaces and packages or IBCs must be clear of oil, ice, or other covering which may reduce friction. Most accidents in transportation are due to improper loading security measures, and carriers or shippers of dangerous goods must ensure appropriate loading security.

6.2. Bulk Transport

A bulk shipment is the transport of a loose, unpacked material (gas, liquid, or solid) in a means of transportation (Fig. 5).

	Bulk transport	
	Liquids, gases	Solids
Road and rail transport	Tank containers Tank trucks Rail tank wagons	Containers, tank containers, bulk containers Tank trucks, silo trucks Sheeted or open, covered vehicles with box-type construction
Marine transport	Ships	
Air transport	Not permitted	

Figure 5. Overview of packagings for bulk transport

Different definitions of the "packed state" of a material apply for the individual carriers.

For shipment by rail and road, dangerous goods are considered packed when they are present in tested and permitted packagings according to Appendix A.5 of ADR [48] or Appendix V RID [49] (max. volume 450 L) or in tested and permitted intermediate bulk containers according to Appendix A.6 ADR or A.VI RID (max. volume 3000 L). For ocean and inland-waterway transport, road and rail vehicles as well as containers already represent a so-called bulk packaging [50].

However, because of the large quantities which are shipped in one packaging, shipment in bulk represents a higher potential of release. Therefore, only less dangerous substances are permitted for this mode of carriage or the packaging for this mode of carriage or the packaging must be designed to be adequately safe (e.g., double wall). The logistic chain is often greatly simplified by bulk shipment, the number of shipments required is greatly reduced, and savings by orders of magnitude are made in packing material.

6.2.1. Bulk Transport of Liquids and Gases

6.2.1.1. Tank Containers [42,51–56]

In the transport of gaseous and liquid goods (as well as granular materials and powders, see also Section 6.2.2), the tank container is one of the most widely used means of transport equipment.

Tank containers are equally suitable for all transport routes (water, road, rail) and therefore, as a multimodal means of transport, offer a variety of possible uses.

These possibilities range from transport via intermediate storage up to long-term storage or stockpiling of goods. With regard to the design and manufacture of the tank, and its testing and operation, regulations (IMDG-Code, based on the UN recommendations) and standards (ISO) valid and accepted worldwide must be observed.

It must be demonstrated by calculation that the tank container withstands the planned loads, i.e., internal and external overpressure and the load which results from the inertia of the tank's contents during movement in transport. The tank, including its attachment fittings, must be able to absorb the following forces at the highest permissible weight of its contents:

1) Twice the total weight in the direction of travel
2) The total weight horizontally perpendicular to the direction of travel (if the direction of travel is not clearly defined, twice the total weight applies in each direction)
3) The total weight vertically upwards
4) Twice the total weight vertically downwards

With these forces acting, there must also be a safety factor of at least 1.5 relative to the elastic limit.

The tank material must be resistant to the cargo or be protected by a resistant lining or coating. Items of equipment on the tank container must be so mounted that during transport or handling they are secure against tearing off or damage. They must guarantee the same safety as the tank, and remain leak-free if the tank container turns over.

Tanks with discharge underneath must be provided with two independent closures in series, the first of the two closures being internal.

Each tank container licensed under the IMDG-Code must be fitted with a pressure-compensating device. This can consist of a melting fuse, a bursting disk, or a pressure-relief valve and must be approved by the responsible authority.

The total amount blown off by the presssure-compensating device from a tank completely surrounded by fire must be sufficient to limit the pressure in the tank to a value of 20 % above the response pressure of the pressure-compensating device.

Each tank container is subject to a type license by the national licensing authority concerned.

Tank containers built according to this license, including their items of equipment, have to be tested before being brought into service and subsequently at regular intervals by experts recognized by the authorities.

In the operation of tank containers the maximum degree of filling must be observed. In tanks of capacity >7500 L, to avoid forces arising from the surging of liquids, a minimum degree of filling of 80 % must be observed, or surgeobstructing arrangements provided.

As a result of the regulations valid and applied worldwide it is guaranteed that tank containers satisfy a uniformly high standard of safety engineering.

6.2.1.2. Tank Trucks and Railroad Tank Cars [51,52,57–64]

In addition to the transport of dangerous materials in tank containers, transport in tank trucks and railroad tank cars is of great importance.

A tank truck is a vehicle with one or more tanks fastened to it for the transport of liquids and gases, as well as powders and granular solids (see also Section 6.2.2).

For *tank trucks,* aside from the constructional and equipment requirements of the actual tank, which are comparable with those of tank containers, the following requirements with regard to the electrical and nonelectrical equipment of the vehicle must be observed.

The rear of the vehicle must be protected against collision by a bumper over the whole width of the tank.

In vehicles that transport liquids with a flash point of <61 °C or flammable gases, the vehicle engine, the exhaust pipes, and the loading pump (if present) must be constructed such that danger to the load resulting from heating or ignition is avoided.

The vehicles' fuel tank must be so arranged that it is protected as well as possible in a collision. If fuel nevertheless escapes, it should be able to run off directly to the ground. Positioning of the fuel tank directly above the exhaust arrangement is not permitted.

In many countries the quality of the vehicles' braking system is subject to testing at regular intervals.

Tank trucks carrying dangerous materials must be equipped with at least one portable fire extinguisher, mainly intended for extinguishing burning tires.

Special requirements must be met by the electrical equipment of a vehicle if liquids with a flash point of <61 °C, combustible gases, or certain explosives are transported:

1) Sufficient dimensions of the cables to avoid overheating
2) Adequate insulation
3) Protection of the leads from impacts, stones, and the heat from the exhaust arrangement
4) Fitting of a battery isolation switch in the driver's cabin or outside on the vehicle, which breaks all circuits in the event of danger

The use of safety valves to avoid unacceptably high pressure in the tank is not uniformly prescribed. In land traffic, it is mainly tightly sealed tanks that are transported.

Arguments against fitting with safety valves are, in particular:

1) Possible operation of the valves (e.g., in tunnels or when driving through built-up areas), whereby the operation is possibly not noticed
2) Possible gumming-up of valves with product and, therefore, failure to function
3) Experience with valves "flaring off" has shown that after a short time the tank and valve material is annealed and loses strength, which can then lead to a tank explosion

Arguments for using safety valves are:

1) In the case of overfilling, overstressing of the tank by hydraulic pressure is avoided
2) The noise of the valve blowing off serves as an indication of danger

As is prescribed in many countries, the drivers of tank trucks should possess a certificate demonstrating that they have taken part in training and have passed an examination on the special requirements to be met in the transport of dangerous materials.

Railroad tank cars consist of an assembly of one or more tanks and the underframe. The basic requirements for the tanks with regard to construction, equipment, and testing are comparable with those for tank containers and tank trucks. Requirements specific to the railroad are set for the underframe, consisting of running carriage, suspension, drawgear, and buffer gear, and for the brakes.

Within the scope of application of ADR and RID, as described in Section 6.2.1.1, both tank trucks and railroad tank cars are subject to type approval by the responsible authorities and regular testing by officially recognized experts.

6.2.1.3. Marine Tankers [65–73]

Whereas the transport of liquids in tank vehicles and tank containers by land is considered to be bulk transport, in shipping, such transport in units regarded as "bulk packaging" is carried out in accordance with the rules for packed dangerous goods. The term bulk cargo is used if unpacked material is taken into the ship's tanks.

The requirements for such ships for the field of ocean shipping are contained in the SOLAS (*International Convention for the Safety of Lifes at Sea*) Conventions and MARPOL (*International Convention for the Prevention of Pollution from Ships*) as well as in the IBC (*International Bulk Chemical Code*) and IGC (*International Gas Carrier Code*) Codes, and the requirements for inland navigation in the ADN and/or ADNR Regulations. Supplementary dimensioning and construction regulations are contained in the appropriate classification and construction rules of internationally recognized classification corporations (e.g., Germanischer Lloyd).

In view of the increasing oil pollution of the oceans and environmental protection in general, measures for raising the standard of safety both in inland-waterway and ocean shipping have been introduced stepwise in the above-mentioned regulations, or have come into force in 1995 for inland shipping and for existing ships in ocean shipping.

Inland Shipping. Inland navigation dispenses with its own list of substances and adopts, with a few additions, that of the land traffic regulations by direct reference in Marginal no. 6002 1a to Appendix A of ADR in its current version. Problems in the transport of packed dangerous goods on transfer from land transport to inland shipping are therefore excluded.

Bulk transport in tankers is permitted only for products listed in Appendix B 2. This list at present contains ca. 270 entries for directly named substances and N.O.S. products of Classes 3 and 8 to which a UN number has been allocated. The list is to be regarded as open, i.e., it can be extended after examination of a request to transport. The time until acceptance is bridged by exemptions regulations limited to one year, according to a precisely defined procedure.

As a result of the requirements included in the substance list for the products and further substance-related regulations in Part II of Appendix B 2 of ADNR, the tanker regulations of the new ADNR are now related to substance rather than type of ship.

In spite of substance-related construction regulations, there are still three types of ship, which however can now be regarded only as a rough classification. The allocation of any product to one of the three types of ship

Type G tanker (gas tanker)
Type C tanker (chemicals tanker)
Type N tanker (normal tanker) is given in the substance list.

The *Type G Ship* is a tanker intended for the transport of gases. Transport is normally carried out in the liquefied state, either under pressure at ambient temperature or at cryogenic temperatures and atmospheric pressure. Partial cooling and transport in pressure tanks is also possible.

For the transport of liquefied gases under pressure, the hull must be constructed in the cargo area as follows:

1) As a double-hull ship with wing walls and double bottom minimum wing-wall width 0.8 m minimum double-bottom height 0.5 m
2) As a single-hull ship with reinforced side-wall structures minimum distance cargo tank – side wall 0.8 m minimum distance cargo tank – bottom 0.6 m
3) Other designs, provided a collision energy of at least 22×10^6 Nm is attained.

The maximum permissible tank capacity depends on the size of the ship (LBH value), but

must in no case exceed 380 m^3 per tank. A minimum number of tanks is no longer prescribed, but proof must be provided of sufficient damage stability when two adjacent compartments spring a leak under precisely defined damage conditions. When the main engine room springs a leak, only the single-compartment status has to be proved.

The *Type C Ship* is a tanker designed for the transport of especially dangerous liquids, such as poisons of Class 6.1 or combustible liquids of Class 3 with additional dangerous properties such as carcinogenicity and toxicity. This is a flush-deck, double-hull ship with wing walls and double bottom, the cargo tanks in the standard version being formed by the hull.

The use of independent cargo tanks also is permissible if they are installed in the inner holds formed by wing walls and double bottoms.

The minimum width of the wing walls is 1 m, but if certain collision reinforcements are carried out it can be reduced to 0.8 m. The average double-bottom height must be 0.7 m, and this dimension must not be less than 0.6 m at any point.

Apart from the requirement of a maximum permissible tank capacity, a maximum cargo tank length must also be observed. The tank capacity depends on the size of the ship, but must not in any case exceed 380 m^3.

Sufficient damage stability when two adjacent compartments spring a leak has to be demonstrated, whereby certain minimum assumptions of the extent of damage in the event of a leak are made.

Here also, the proof of single-compartment status is sufficient for the main engine room.

There is no need to assume damage to the inner longitudinal bulkhead of the wing walls; likewise the crown of the tank is to be regarded as undamaged, provided the double-bottom height is at least 0.6 m.

Type C double-hull ships with closed tanks must be designed for pressure ratings that rule out a response of the safety valves during the transport of toxic and carcinogenic substances.

A reduction of the safety valve setting can be obtained by additional sprinkling, provided pressure sensors ensure that the sprinkler system comes into operation early enough.

The *Type N Ship* is a conventional tanker for the transport of liquids. These are usually single-hull ships of various pressure ratings for the transport of combustible liquids of Class 3 (comparable with the former tanker Types II to V, but with altered classification criteria).

Acid and alkali tankers for the transport of goods of Class 8 are of this type, even though they are mostly double-hull ships or ships with inserted tanks. With these ships special attention must be paid to material compatibility and it should be noted that acids and alkalis generally have a relatively high density.

Particular distances of tanks from the ship's hull or from the bottom for safety reasons are not required for these ships, nor is proof of damage stability. Sufficient intact stability is, however, prescribed (see Marginal no. 331 214).

In addition to the maximum permissible tank capacity (environmental protection), a maximum cargo tank length must be complied with. The former requirement for a minimum number of tanks has been dispensed with.

There are few restrictions regarding the construction of Type N Ships, so the following alternatives exist, unless they are restricted by the invidual requirements of the substance list:

1) Flush-deck, single hull ship
2) Single-hull ship with trunk
3) Flush-deck, double-hull ship
4) Double-hull ship with trunk
5) Ship with inserted tanks
6) Ship with inserted pressure tanks

The basic safety requirements are the same for all Type N Ships. However, Marginal no. 210 287 (Appendix B 2, Chap. 1) exempts tankers of Type N, open and with and without flame arresters from the observance of a number of requirements.

Thus, the requirements for tankers of pressure rating 0, i.e., those having tanks with an open venting system, are effectively not increased. Only the number of products that can be transported in a tanker with flame arrester has been restricted.

A simplified comparison of the Type N Ships with the former tanker Types I to V is shown in Figure 6.

Ocean shipping has its own list of substances in the IMDG-Code, which in spite of having the same classification, deviates from that of land transport in the ADR, although both

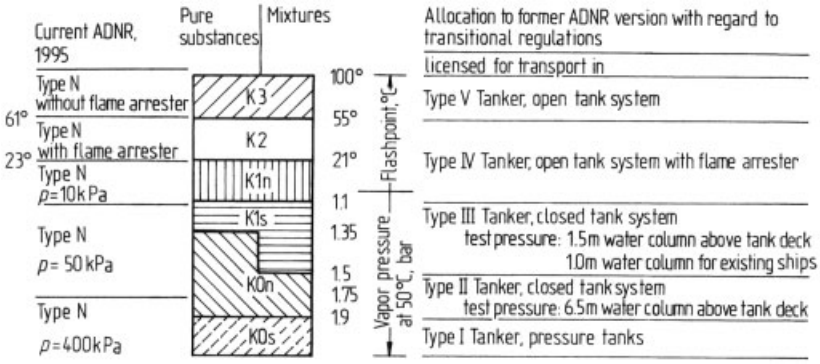

Figure 6. Simplified comparison for Type N Ships with former Type I to V Tankers

originate from the list of substances in the UN recommendations.

The basic regulations for the bulk transport of liquids, as far as dangerous goods are concerned, can be found in Chapter VII of SOLAS and in the MARPOL Convention. These are supplemented by the IBC and the IGC Code.

MARPOL 73/78 is the International Convention of 1973 for the *Prevention of Pollution from Ships* in the version of the protocol of 1978. Appendix 1 of MARPOL 73/78 contains the rules for the prevention of oil pollution by ships and came into force, in part retroactively, on 2 October 1983.

"Oil" here means all petroleum products, including crude oil, fuel, oil sludge, oil residues, and refinery products such as fuel oil and gasoline. "Oil tankers" are defined as ships that transport oil as bulk cargo.

At the 31st MEPC committee meeting in July 1991, the text of the new Rule 13 F for oil tankers was formulated and adopted. It applies to new oil tankers of >600 t load whose construction contract was concluded after 6 July 1993.

Rule 13 F requires for oil tankers of ≥5000 t capacity that the whole cargo tank length is protected as follows by ballast tanks or holds that are neither cargo nor fuel tanks:

1) By wing tanks or spaces over the whole depth to the deck of the ship or from the double bottom to the topmost deck with a width in meters of

$$w = 0.5 + \mathrm{DW}/2000,$$

where DW is the load of the ship in tonnes deadweight (tdw). The minimum value for $w = 1$ m, and the maximum required value is w=2.0 m.

2) By a double bottom with a height in meters of

$$h = B/15,$$

where the minimum value for $h = 1.0$ m and the maximum required value is $h = 2.0$ m (B = breadth of ship).

3) In addition, the total capacity of the ballast tanks must be sufficient for a safe voyage in ballast with the screw fully immersed.

For oil tankers of <5000 t load, equivalent solutions are provided.

As a precursor of MARPOL Rule 13 F, Rule 13 E for crude oil tankers above 20 000 tdw and for product tankers above 30 000 tdw, in effect from 1 June 1979, required a partial double hull by having segregated ballast tanks in a protective arrangement. "Segregated" means that these ballast tanks must be completely separated from the oil cargo and fuel systems; i.e., they can only be used for ballast water. This partial double hull can be formed by a double bottom and/or wing tanks.

The degree of protection J was 0.45 for tankers of 20 000 tdw and 0.3 for tankers of 200 000 tdw. Intermediate values are found by linear interpolation. For tankers of >200 000 tdw J could be reduced to as low as 0.2 according to Rule 23/24, depending on the ratio of the hypothetical to the permissible oil outflow.

In Rule 13 G measures are provided for existing tankers for the prevention of oil pollution in cases of collision and stranding.

Various measures are provided for crude oil tankers of ≥20 000 tdw and product tankers of ≥30 000 tdw for raising the safety standard of existing tanker fleets. These measures can be summarized as follows:

1) Stricter inspections. In addition keeping a complete record of the inspection reports, which is available to the responsible authorities of each of the signatory states.
2) Conversion to a double-hull ship or laying-up after a ship's lifetime of 25 a. For ships that fully satisfy Appendix I of MARPOL 73/78, i.e., those for which the construction contract was concluded after 1 June 1979 (so-called MARPOL tankers), a period of 30 a is laid down.

The requirements of Rule 13 G came into force on 6 July 1995.

It can be gathered from the above that according to Rule 13 F new oil tankers in ocean shipping with a load >5000 t must be built over the whole load range with wing tanks and double bottom, and for those with a load <5000 t equivalent regulations must be observed. However, according to Rule 13 G, for the existing tanker fleet a stricter scope of inspection is required only for crude oil tankers of ≥20 000 tdw and product tankers of ≥30 000 tdw, together with conversion to double-hull ships or laying-up after 25 or 30 years, respectively.

For existing smaller tankers, corresponding precautionary measures have not been agreed upon, so that a definite final date for the existence of small "single-hull tankers" depends on their lifetime, since new ones may not be built.

Apart from the crude oil tankers and product tankers that correspond to the Type N Tankers in inland shipping, ocean shipping has according to Chapter VII of the SOLAS Convention declared the International Bulk Chemical (IBC) Code and the International Gas Carrier (IGC) Code to be binding regulations.

The IBC Code contains material-related regulations for the transport of dangerous liquid chemicals in bulk, allocated to the three types of ship:

1) A *Type 1 Ship* is a chemical tanker intended for the transport of materials that constitute a very great danger with regard to environment and safety and require the highest possible level of safety precautions against the escape of cargo
2) A *Type 2 Ship* is a chemical tanker intended for the transport of materials that constitute a great danger with regard to environment and safety and require considerable safety precautions against the escape of cargo
3) A *Type 3 Ship* is a chemical tanker intended for the transport of materials that constitute a fairly great danger with regard to environment and safety and require a certain measure of safety precautions to improve the buoyancy under damage conditions

Arrangement of the cargo tanks:

Type 1 Ships:	*B*/5 or 11.5 m from the side (the smaller value is required)
	B/15 or 6 m from the bottom (the smaller value is required)
Type 2 Ships:	*B*/15 or 6 m from the bottom (the smaller value is required)
	760 mm from the side
Type 3 Ships:	no requirements (i.e., Type 3 Ships are single-hull ships)

Apart from intact stability, damage stability and buoyancy in the leaking state under specified damage conditions must be demonstrated.

The regulations relevant to substances can be obtained from the substance list, comprising about 800 products, as well as from the special chapters: "Special Requirements," "Additonal Measures for Protection of the Marine Environment," and "Operating Instructions."

The IGC Code governs the sea transport of gases liquefied under high pressure and by cooling. Depending on the potential danger originating from the gas, one of the following specific types of ship is prescribed: Type 1 G Ship, Type 2 G Ship, Type 2 PG Ship, or Type 3 G Ship.

A Type 1 G Ship is a liquefied gas tanker intended for the transport of materials with the highest potential danger, while Type 2 G/2 PG and Type 3 G Ships are provided for the transport of materials with lower potential danger. Accordingly, a Type 1 G ship should be able to survive the severest assumed damage and its cargo tanks have to be arranged at the greatest prescribed distance from the side plating.

If a ship is intended to transport more than one substance of those listed in Section 19 of the IGC Code, the assumptions about damage must correspond to the substance for which the highest demands are made on the type of ship.

The requirements for the arrangement of the individual cargo tanks, on the other hand, refer to the substance actually to be transported in each case.

The Code contains precise specifications with reference to the arrangement of the cargo tanks.

The damage stability and the buoyancy under prescribed damage conditions must be demonstrated, in addition to the intact stability.

6.2.2. Solid Bulk Cargo

Bulk cargo is a loose material (other than liquid or gas) in pourable form [74].

It is characteristic of bulk material that the components or particles are disordered and mutually displaceable, so that the material as a whole has a variable shape and is fluid or pourable. Gravitational force can therefore be exploited for moving bulk material.

Characteristic physical quantities of bulk material are cohesion and adhesion, bulk density, angle of repose, and grain shape and grain size d:

Dust	$d < 0.06$ mm
Fine	$0.5 < d < 10$ mm
Coarse grained	$10 < d < 60$ mm
Medium	$60 < d < 160$ mm
Lump	$d > 16$ mm

These material quantities – together with the hazard properties of the bulk material such as toxicity, gas evolution, corrosivity, infectiousness, flammability, and explosiveness – form the basis for choosing a suitable packaging.

6.2.2.1. Solid Bulk in Road and Rail Transport

Definition. In road and rail transport, the shipment of dangerous material in bulk is defined as "the shipment of a solid material without packaging" [75,76]. This definition is not entirely correct, since by using the term "material" it excludes an important area of bulk cargo shipment: the disposal of uncleaned packages, contaminated objects, and equipment. In general therefore, one should refer to the shipment of solid goods without packaging.

Permitted Goods. The road and rail dangerous goods regulations name three modes of shipment of goods: shipment in bulk, shipment in containers, and shipment in tanks.

This system is illogical and contradicts the definition of shipment in bulk. According to ADR/RID the "bulk" mode of shipment is permitted only for certain materials [76]:

Class	Number	Materials
4.1	4 c)	wastes
	6 c)	all materials and wastes
	11 c)	all materials and wastes
	12 c)	all materials and wastes
	13 c)	all materials and wastes
	14 c)	all materials and wastes
4.2	1 c)	all materials
	2 c)	all materials
	3 c)	all materials
	12 c)	2793 iron metal swarf
	16 c)	1376 iron oxide, used; iron oxide sponge, used
4.3	11 c)	all materials
	12 b)	1405 calcium silicide pieces
	12 c)	all materials
	13 b)	3170 aluminum sweepings
	13 c)	all materials
	14 c)	all materials
	15 c)	all materials
	17 b)	all materials
	20 c)	all materials
5.1	11	all materials and wastes
	12	all materials and wastes
	13	all materials and wastes
	16	all materials and wastes
	18	all materials and wastes
	19	all materials and wastes
	21	all materials and wastes
	22 c)	all materials and wastes
6.1	44 b)	all materials and wastes
	60 c)	all materials and wastes
	63 c)	all materials and wastes
	all in c)	all wastes
6.2	1	all
	2	all
	3	all
	5	all
	9	all
8	1 b	sulfuric-acid-containing lead sludge
	23	all materials and wastes
	all in c)	all wastes
9	4 c)	2211 foamable polymer beads
All		empty (uncleaned) packagings

For all unlisted materials it is forbidden. The selection of materials is not subject to any system but rather is governed by need. Owing to the assumed low safety level of the packings in question, however, the less dangerous materials (c materials) predominate. For some special

areas (e.g., spent accumulators, contaminated soil), there are exceptional national regulations or regulations specific to federal states or special agreements with individual ADR/RID treaty states.

A considerably greater number of bulk materials is permitted according to ADR/RID for shipment in tanks, since relevant constructional, equipment, and testing requirements guarantee a high safety level of the packaging "tank" (see Sections 6.2.1.1 and 6.2.1.2).

Means of Transport and Requirements. For the shipment of bulk material by land, the following means of transport are mainly used (Fig. 7):

1) Open, sheeted, or covered vehicles with box-type construction (dumper trucks)
2) Tank trucks with pressure discharge, or silo vehicles
3) Containers/bulk containers according to ISO 1496/ Part 4 [77]; so-called logistical containers, e.g., put-down containers, according to DIN 30 723 [78] and roller containers according to DIN 30 722 [79]; tank containers

The bodywork of the vehicles is not subject to test requirements. According to the definitions in ADR/RID, a covered vehicle is simply a vehicle whose body can be closed. A sheeted vehicle is one that is provided with a canvas cover to protect the load. The design of the bodywork, like that of containers, remains largely unrestricted in respect of mechanical stability and watertightness.

If materials permitted according to ADR/RID for shipment in bulk are shipped in tank trucks or tank containers, the regulations for construction, equipment and operation of tanks are not applicable [80], i.e., for tank shipment also there are no minimum requirements for stability, watertightness, and equipment.

For the shipment of a number of flammable or spontaneously flammable materials a metallic vehicle construction, and in some cases the use of impermeable or flame-resistant sheeting is also required.

The vehicle bodywork or container for spontaneously flammable materials or materials that evolve flammable gases on contact with water must be capable of airtight closure. For ignitable materials (Class 4.1) and nauseating materials (Class 6.2), sufficient ventilation is prescribed. In the shipment of materials with igniting (oxidizing) effect (Class 5.1), contact with wood or other combustible materials must be prevented.

Figure 7. Bulk container
Top: 12-m silo tank container for the shipment of toxic dust; Bottom: 4-m^3 silo tank for shipment of toxic lead compound

In various fields, particularly in the disposal of industrial wastes, there is a requirement for packages for the collection and shipment of waste, which can guarantee safe disposal of even dangerous wastes, sometimes as heterogeneous mixtures, in large amounts ($>3\ m^3$), and whose construction is adapted to the logistical requirements of an optimum waste-disposal chain (e.g., large openings for filling and emptying).

There have already been a number of special developments whose design is based partly on requirements from the tank field and partly on the regulations for IBCs, as well as additional material requirements (Fig. 8). For these shipment concepts, a special license from the responsible authority is necessary.

Figure 8. Special bulk container for shipment of hazardous waste (Courtesy of HIM)

Starting from the assessment that the current regulations for bulk shipment in land transport do not correspond to the state of the art and do not take sufficient account of the current requirements, a rearrangement of this shipment field is being contemplated that should lead to systematic and clear regulations and objective test requirements.

6.2.2.2. Solid Bulk in Marine Transport

Definition. The regulation for marine and inland-waterway transport define bulk solids as "goods, other than liquids and gases, that can be loaded directly without any intermediate packing into the hold of a ship." Goods are included that are loaded in a barge carrier's lighter on a barge carrier. In addition, the term "bulk package," which refers to packaging (containers, road or rail vehicles) that contain bulk cargo, has been introduced in the field of marine transport [50].

Permitted Materials, Requirements. Solid materials in bulk packages may be shipped in marine transport if, according to the relevant substance page of the IMDG-Code, bulk shipment is permissible. The permissible range of materials corresponds approximately to that in land transport. The materials must be shipped in closed containers or closed road or rail vehicles with superstructures or freight car bodies of metal (dust-tight wooden floors are permissible) or in mobile tanks. The relevant land transport regulations are also to be observed.

For some materials there are also special requirements for the bulk packages. Solid materials that are shipped in bulk in the cargo hold of a ship and from which, because of their chemical or physical properties, a danger originates during transportation are listed in Appendix B of the Code on the Safe Handling of Bulk Cargoes (BC Code).

Some of the materials listed therein represent no substantial danger if they are transported in packagings. However, shipped in bulk, they represent a hazard. These substances include:

1) Aluminum wastes
2) Fly ash
3) Activated carbon, coal, coke
4) Direct-reduced iron, iron phosphide
5) Fluorspar
6) Quick lime
7) Magnesium
8) Pitch
9) Sawdust, wood shavings, wood pellets
10) Silicomanganese
11) Vanadium ore

For materials classified as dangerous goods, the corresponding substance pages of the IMDG Code contain a reference to the BC Code. Before each shipment in bulk of an unlisted material, comprehensive information on the physical and chemical properties of the material can if necessary be obtained by request to the responsible authority.

The prescribed safety measures gathered in the "instructions for the safe handling of bulk cargoes in transportation by ocean ships" are adequate to match the potential hazard of bulk cargoes of dangerous material on ships.

6.3. Cargo Storage

For the storage of goods in packages, logistic considerations, the type of store, and safety requirements all have to be harmonized.

The logistic plan is determined by the type and number of goods, the packing, the storage units (pallet units), the turnover rate, and the type of carrier.

The kinds of storage that must be distinguished are block storage, for large lots of stock units of the same type, and shelf storage, for the storage of single packages and pallet units.

In the design of the store, logistic and safety aspects have to be taken into account:

1) Unroofed open-air storage consists of a paved area, impermeable to liquids, and sufficiently firm. Products can be stored here that are insensitive to temperature and have weather-resistant packing. Individual storage zones are separated by safe distances.
2) In roofed open-air storage the stored goods are shielded from the direct effects of the weather by a self-contained roof structure, but the store has no continuous external walls. The stored goods must be insensitive to temperature. Individual storage zones are separated by safe distances or by walls.
3) In storage buildings, a distinction is made between single-level, multilevel, and racked stores. Products with different safety requirements are accommodated in different store rooms. Storage zones are separated by fire-resistant walls or ceilings. The store is suitably divided into loading zone, handling area, and stocks on hand.

In siting the store facility, sufficient distances must be maintained from residential areas, busy traffic routes, and special protected areas such as drinking water extraction areas. Areas with risk of flood, earthquake, landslide, and avalanche must be avoided. The infrastructure must include good traffic connections, good accessibility for the fire brigade, and a supply of water for fire fighting.

The safety plan for a store must take into account the potential dangers to employees, the neighborhood, and the environment. Control measures serve to avoid a danger arising, to prevent the danger spreading, and to fight the danger. Rules and prohibitions governing storage together of incompatible substances must be observed.

The release of dangerous substances, especially toxic and combustible liquids and gases, is prevented by the choice of corrosion-resistant, stable packing drums, by prohibition of inadmissible overstacking, and by safety devices to prevent falling from shelves.

Escaped product is retained in durable collecting troughs or in the storeroom. The collecting devices must be liquid-tight and product-resistant. With readily volatile liquids and gases, it may be necessary to close openings to the building and switch off the artificial ventilation.

Measures for explosion protection and ventilation are used for the avoidance of fires and explosions, especially with combustible liquids and gases. Ignition sources must be avoided on principle.

The spreading of fire is hindered by constructional arrangements (fireproof walls, ceilings, and load-bearing parts of the building) and by specific store organization (separation of the storage zones by walls and spacings, buffer zones with nonflammable materials). In fairly small storerooms, the spreading of fire can also be hindered by a gas-tight mode of construction and automatic closure of openings (ventilation-controlled fire).

The fire is fought initially by the store personnel, using hand fire extinguishers and wall hydrants. The fire alarm is given by press-button alarm, by telephone, or by automatic fire alarm installations connected to the responsible fire department. The nature, equipment, and speed of the fire brigade must match the danger potential. If the danger potential is considerable, it is necessary to equip the store with automatic extinguishers.

The fire protection plan, including selection of the extinguishing agent, is drawn up with the local fire department. The nature of the fire protection methods also influences the arrangements for retaining contaminated fire-fighting water.

In addition to constructional and technical measures, organizational rules are necessary for the safe and economic operation of a store. In particular:

1) A storage plan and a stock list; both used for rapid information and orientation in the event of danger
2) An internal alarm and accident prevention plan
3) Written instructions for the store personnel on handling dangerous substances, and on rules of conduct in the event of danger

6.4. Bulk Storage

6.4.1. Storage Concept [82–95]

When storage concepts are being developed, the dangerous properties of substances must be taken into account. The optimum amount stored for a store of a given scale follows from production and sales considerations. This can raise an area of conflict between reliability of supply and profitability, on the one hand, and the protection requirements of employees and third parties on the other. Protection of third parties includes neighborhood and environmental protection.

The following forms of stationary tank are customary for the storage of bulk liquids: horizontal and vertical cylindrical tanks, spherical tanks, and flat-bottomed cylindrical tanks in a collecting space. The following types of installation can be considered: above-ground: in the open or in a containment; underground: corrosion-protected; in a collecting space or double-walled or soil- covered with an open end.

The following substance-related risk factors exist: amount stored, pressure, energy potential (product of pressure and capacity), temperature, explosion danger, combustibility, hazard to health, chemical reactivity (e.g., ethylene oxide), corrosion risk (e.g., hydrogen-induced stress corrosion cracking by ammonia in ferritic steels), and danger to water sources. Further intrinsic risk factors arise from the design of the tank system and from processes in which the tank takes place.

The following environment-related risk factors are important: site conditions, fire loads, ignition sources, storage together (layout), foundation (e.g., in mining or earthquake districts), traffic, natural risk factors (e.g., flood danger) and artificial ones. Operational risk factors include human errors in filling, emptying, sampling, cleaning, and maintenance, as well as leaks during operation.

The inclusion of a store in an existing infrastructure reduces risk, e.g., with regard to the reliable supply of energy and water as well as to the immediate engagement of a works fire department.

Changes to existing safety rules must follow the same protective goals. The goals can also be met by other appropriate methods, with consideration of the balance between cost and safety gain.

In this connection, control of disturbances by the third barrier (organizational measures) and the fourth (limitation of consequential damage) must be adequately evaluated when designing the first barrier (tight enclosure) and the second (sufficient safe drainage of any leakage). In works with a well-developed infrastructure, especially in the chemical, petrochemical, and petroleum industries, this third barrier is always readily available.

Principles of a safety plan comprise:

1) Keeping stored amounts small: reducing total amounts stored to the quantity necessary for reliable supply.
2) Lowering the hazard potentials: storage in chemically stabilized or dilute form, mixed with inert carrier material, in solvents, or at low temperature.
3) Distribution: amounts of substance distributed between discrete sections of the store, separated by fire walls.
4) Storing substances separately: this requires that dangerous properties (explosiveness, combustibility, hazard to health) be itemized. For combustible substances it is important to know whether they are readily flammable or even spontaneously flammable; such substances should be stored separately. It is advisable to classify substances into groups according to the transport regulations, extended to include chronic toxicity, carcinogenicity, crop-damaging and genetically modifying properties, and potential danger to water sources. Combustible substances should not be stored together with igniting or spontaneously flammable substances, or with toxic substances, or gases that are compressed, liquefied, or dissolved under pressure.

Avoiding Environmentally Dependent Hazards. It is important to check the site; eliminate fire loads; avoid ignition sources where combustible substances are concerned; separate substances of different danger classes; secure foundations (taking note of soil surveys, earthquake zones); protect against mechanical damage; provide protective distances (e.g., ex-

plosion protection) and safe distances (e.g., from hospitals and schools).

Optimizing Organizational Protective Arrangements. The following requirements must be set: operating instructions (including avoidance of human errors, in the initial filling and in the event of trouble); qualified personnel; regular training; cleaning and maintenance permits; safe infrastructure (energy supplies, water reservoir, fire protection, waste disposal); and practices of alarm and accident prevention plans.

Avoiding Hazard Activation: Maintaining Primary Safety (First Safety Barrier). The leak-tightness of the plant (tank, piping, equipment, feeding devices, seals) must be continuously monitored. The following protective goals are important:

– Equipment to be designed for pressure, temperature, and other stresses
– Compatibility to be preserved between equipment materials and the stored substance
– Confusion of substances to be avoided (control of substances delivered)
– Materials to be sufficiently resistant to or protected from weathering or other external influences
– Protective equipment to be reliable (redundant, e.g., over-filling protective devices LS^+A^+ and level indicators LIA^+), fail-safe, or self-monitoring
– Alarms and monitors to be provided
– Automatic leak-testing to be provided at separable connections
– Protection to be provided against heat, fire, and cold
– Adequate grounding to be provided
– Protection to be provided against mechanical action from outside, e.g., guide planks
– Arrangement of the installations to be clear and understandable
– Filling systems to be situated outside the storage area
– Human errors to be reduced by organizational measures (operating instructions), e.g., rules for acknowledgment action taken, and by technical measures (e.g., over-filling safety devices with automatic disconnection LS^+A^+)
– Parts subject to wear (e.g., hoses) to be replaced in good time
– Safety equipment (e.g., over-filling safety devices) to be given regular operational tests with appropriate documentation
– Regular leakage checks with appropriate documentation to be carried out
– Periodic tests to be carried out on vessels, lines, hoses, devices for avoiding electrostatic charging, and for protection from lightning and fire

Limiting the Effects of Malfunctions (Second Safety Barrier). Malfunctions can be detected early by regular monitoring (e.g., patrols). The following protective measures can be suitable:

In store rooms (housing), regular monitoring to be supplemented by automatic early warning equipment (e.g., gas warning devices), optionally with suction into a destruction system.

Components to be designed according to the "equal safety" principle (structures of the same strength) and "leak before break": fracture of a pipe connection and catastrophic failure is then impossible, and need not be included in the considerations.

Emergency shutdown systems with quick-acting gate valves to be applied when transport tanks roll away, and optionally automatically effective in case of fire.

Emergency nitrogen and power supplies to be available if the fail-safe principle is not used.

Pipe-break safety devices or quick-acting gate valves to be incorporated in lines that convey liquid.

Quick-separation systems to be used in ship reloading.

Full-hose systems to be used at reloading or filling stations.

In connection with filling with hoses or articulated arms: quick-acting fittings to be used on both sides, in above-ground installation of tanks for combustible substances in the liquid state: good ventilation to be applied; soil areas below the tanks to be liquid-tight and drained by a safe route.

Tanks to be installed in accordance with the properties of substances, e.g., very toxic gases in the liquid state in collecting spaces, double-walled tanks with leak monitoring, or as single-walled accommodation with suction; very-low-temperature storage with collecting space for

toxic and combustible substances; as a rule, above-ground installation in the open air is best with regard to safety and cost, since the tanks can be monitored and maintained in the best possible way throughout their life.

Electrical and mechanical explosion protection to be provided.

In the case of danger to water sources, arrangements to be made for the collection of leaks (e.g., collecting spaces or double-walled tanks with leak monitoring).

No gully holes and ignition sources to be present in the protected zone, especially for separable connections.

Water sprinkling to be provided against the effect of solar radiation.

Safety valves to have a safe discharge line.

Possibility of covering the collecting space with foam or, in the case of combustible substances, filling the collecting space, e.g., with gravel.

For confined liquid volumes, e.g., in pipelines, relief valves to be present, in the case of toxic substances with optional downstream flash vessels and upstream bursting disks and buffer vessels.

Fire protection plan to be coordinated, e.g., sprinkling or flooding with water; water sprinkling by the works fire department, but only when the heat is exclusively radiant; protective gap; protective wall; fire insulation; insulation or soil cover. For installation above ground with water sprinkling, a protected extinguishing agent supply and fire protection equipment to be provided; for remote sites without continuous monitoring, optional water sprinkling or flooding with fusible fire safety device and automatic response; amounts of water sprinkling must be laid down as a function of tank size and type; for large spherical tanks ($>1000\,\mathrm{m}^3$), sprinkling densities of $\geq 100\ \mathrm{L\,m^{-2}\,h^{-1}}$ are sufficient.

Measures to be provided against fire under a tank and spontaneous combustion.

Personal safety equipment to be provided.

Alarm and accident prevention plans to be drawn up.

The fire protection plan in large works complexes with corresponding infrastructure includes: active explosion protection, qualified operators always present, specially trained works fire department always available, access to the basic pieces of equipment by agreement with the fire department for cooling and, if the need arises, for extinguishing, stock of water always available, maintenance, preventive maintenance, protected energy supplies.

Before commissioning, the safety plan of the plant is to be checked (plant monitoring). The important aspects are the following: substance data, piping and instrumentation flow charts, space-assignment plan, notice of approval, and, if occasion arises, safety analysis and operating instructions.

The acceptance inspection is concerned with operation according to the regulations and possible disturbance of this operation.

The conformity test is based on the observance of the applicable safety regulations. It assumes that there are appropriate vessels, piping, fittings, pressure generators, warning, alarm, and safety equipment, as well as seals, in accordance with the documentation used for checking of manufacturers' specifications as well as the piping and instrumentation flow charts. In the space-assignment test, the following considerations have to be assessed: fire protection, explosion protection including electrostatic charging, health protection, water protection, and infrastructure.

Requirements and design of the safety equipment are based on the process-engineering relationships, the characteristic quantities for safety engineering (properties of the substances), and the local conditions, e.g., safe discharge.

An operational test is to be made of the warning, alarm, and safety equipment. Appropriate documentation on the plant, based on the tests carried out, is to be drawn up.

The special conditions to be taken into account for individual cases, e.g., installation in a water protection area or in the neighborhood of sensitive equipment outside the works, must also be subjected to appraisal in the storage plan. In this connection, the fourth safety barrier is of decisive importance.

In some countries, detailed regulations exist on the handling and storage of dangerous substances. In Germany, the *Technische Regeln für Gefahrstoffe* (TRGS) incorporate the guidelines of the European Community, especially on the classification, packing, and labeling of dangerous industrial substances according to the level of technology employed.

6.4.2. Tanks [91,92,96–102]

In almost every branch of industry tanks are used worldwide for the storage of a variety of liquids, while bulk materials and silage fodder are stored in silos. Tanks are mostly made of ferritic steels, but also of high-grade steels, nonferrous metals, and plastics, e.g., glass-reinforced plastics (GRP). For economic reasons, ferritic sheets with cladding of high-grade steel or nonferrous metal are frequently used in tank construction. Good alternatives are ferritic steel tanks with plastic internal coatings, rubber linings, enameling, and internal coverings applied by means of vacuum (between covering and steel) to the supporting walls. Depending on the construction, it is normal to refer to flat-bottomed tanks (up to more than 100 m in diameter), cubic tanks, cylindrical tanks, double-walled tanks, double-bottomed tanks, and spherical tanks (spherical containers). Another characteristic is the permissible over-pressure, e.g., quasi-pressureless tanks ($p < 10$ kPa) detonation-resistant tanks (these may deform plastically in an explosion, but not develop leaks), and tanks with internal over-pressure ($p > 10$ kPa). For the storage of large quantities of gases liquefied under high pressure, spherical tanks (spherical containers) are preferable, which rest on legs, columns, or eggcup-shaped bases. Flat-bottomed tanks are mostly made with a fixed roof; above ca. 60 m diameter they are built as floating-roof tanks for economic reasons. For the storage of gases liquefied at low temperature (e.g., ammonia, oxygen, nitrogen, argon), the tanks are built with ca. 1.5 m thick special insulation and for permissible operating temperatures down to $-200\,^{\circ}$C or below. For substances that are not pumpable at ambient temperature, the tanks are provided with direct or indirect heating. The usual tank equipment includes: liquid-level indicators, over-filling safety devices, air-inlet and air-removal connections, gas displacement connections, filling and withdrawal connections with shut-offs. For the storage of combustible liquids with flash point $< 55\,^{\circ}$C, the tanks must be earthed and have flame arrestors in the respirator lines. In the case of flat-bottomed tanks for combustible liquids, the weld seam between cylinder and roof must be fabricated as a rated-rupture seam, so that with an inadmissible over-pressure only the roof tears off, and the stored liquid remains in the tank. The tanks are erected, preferably above ground (simple checking of corrosion damage and gas-tightness from outside as well as better accessibility during cleaning, repair, and alteration work), standing vertically or horizontally on stable concrete foundations. In the case of tanks erected above ground, catwalks, stairs, piping, and similar equipment must be so arranged that if there are changes of length as a result of temperature changes or tank settlements, they exert no inadmissible additional forces on the tanks.

Tanks for combustible and water-polluting liquids must be erected within bund walls. Flat-bottomed tanks for water-polluting liquids must be so erected or equipped that leakages in the bottom of the tank can be detected reliably and quickly. This can be achieved, e.g., by:

1) Erection on continuous concrete foundations with plastic or metal foil separating layers or the detector of a leak-detection system between the tank bottoms and the foundations
2) Erection on strip footings
3) Double-bottomed tanks with an inspection hole or leak-detection device

In Germany, steel tanks for the underground storage of water-polluting liquids must be double-walled and the leak-tightness of the interspace must be monitored by a suitable leak-detection device. In the case of underground steel tanks, the use of a cathodic corrosion protection system is advisable. For reasons of explosion-proofness and environmental protection, newly erected tanks are increasingly operated closed, i.e., with no air admission to the atmosphere and with inert gas blanketing (e.g., nitrogen).

Depending on the properties and the amount of stored liquid, the tanks are subject in many countries to national regulations for construction and equipment as well as for the environment, in particular for the protection of water sources. The detailed requirements regarding tank construction, equipment, erection, and testing contained in these regulations and in technical rules and standards associated with them are on the whole comparable as far as above-ground storage is concerned.

Tanks in operation must be tested at regular intervals in accordance with the national regulations by experts recognized by the authorities.

These experts are personally responsible for the operational safety of the tanks tested. Testing intervals are determined by the requirements of the authorities, or by the experience of the operator or builder. Testing must cover the whole tank structure, including all associated equipment. After being taken out of service, the tanks must be cleaned in a professional and environmentally acceptable way, and disposed of likewise.

7. Pipelines

7.1. Types of Pipeline

Pipelines transport and distribute substances in bulk, and are used for a variety of purposes in industry and technology. Their design specifications, layout, and operation are therefore covered by a range of technical and legal codes and standards.

Pipelines can be used as transport systems between two points (line systems), or as network systems for transport and distribution. This applies for long-distance pipelines and intra-plant systems. Typical applications for pipeline systems are for oil, gas, or water transport over long distances. Typical networks are intra- company distribution systems, communal water supply or sewerage systems, and airport fueling lines.

Pipeline systems should always be viewed as a whole. This holistic view of the transport line or network system should include the plant and facilities involved at both ends of the pipeline (Fig. 9), such as tanker ships, rail tank wagons, or road tank lorries. At the input end this may also include the source (e.g., the oil well) where this feeds into the pipeline, either directly or via a collecting tank. At the output end there may be tankage and transport plants, as well as user facilities (filling and transfer plant). The pipeline also includes all fittings, pumps valves, etc. along its length.

Pipeline systems can be laid overground or underground, or as a mixture of both. They are used to supply substances, or to remove unwanted substances (notably sewage). It is important to distinguish between pressure pipelines and gravity pipelines. Gravity pipelines are usually used only for the transport of water and sewage.

Pipelines are used to transport and distribute fluids such as oil and its derivatives, water, sewage, brine, and fluids for production, such as natural gas, hydrogen, acetylene, as well as suspensions of fine-grained solids such as cement, coal, ore, or sulfur. It is important to distinguish between flammable and nonflammable substances, since construction and operating requirements differ considerably between these.

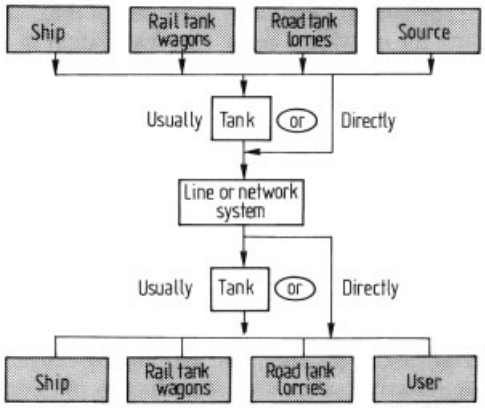

Figure 9. Components of a pipeline system

The numerous possibilities are covered by extensive technical codes, standards, and laws, and the relevant national requirements must be taken into consideration when applying for operating licenses.

7.2. Requirements for Pipeline Systems

Pipeline systems must meet two basic requirements:

1) They must be designed for continuous operation, to ensure supply or removal of the substances involved (which is in the interests of the operators).
2) To obtain permission to operate a pipeline, whether for long-distance transport or for intra-plant operation, it must be shown that the construction and mode of operation exclude the possibility of uncontrolled release of substances into the environment during normal operation, and in the event of damage or disturbances. This requires a redundant safety system, both to protect public safety

(avoiding fires and explosions) and to pollution of soil, ground and surface waters.

The redundant safety system consists of two barriers. The primary barrier is that directly enclosing the substances, with all safety and auxiliary fittings. The secondary barrier serves to contain substances released in the event of damage or disturbance, and also any unavoidable leak and drip losses. This barrier is generally designed as a leakproof system with appropriate safety features and warning systems. Special attention needs to be paid to the transfer points at both ends of the pipeline system, since these represent weak points. This is particularly important for systems involving the discharging or loading of ships.

Extensive technical and operating requirements have been introduced, aimed at preventing unacceptable environmental damage. It is important to establish the substances which are to be transported by a system and the volumes involved, since this influences the choice of materials for the primary and secondary barriers, as well as the range of possible environmental pollution.

The toxic properties of transported substances are of vital importance. The greater the potential threats posed by these substances to the environment, the more care needs to be taken in the design and operation of the pipeline.

In addition to direct toxic effects of substances, whether to humans or more generally in the environment, other environmental effects may also need to be taken into account. Natural gas, for example, is toxic only to the extent that it carries traces of substances such as the carcinogen benzene. However, if released underground it can lead to the displacement of oxygen, with adverse effects on roots and soil microorganisms. Accompanying changes in soil conditions such as pH and redox potentials can also lead to the mobilization of heavy metals.

7.3. Design and Operation

7.3.1. Routing

A well-considered route for a pipeline can go a long way toward meeting the safety requirements for the system. This applies to both external influences on the pipeline and the effects of the pipeline on its surroundings. It is important, for example, to avoid heavily populated areas, areas of importance for water resource management, and mining areas. Where this is not possible special precautions must be taken.

Where roads or railways cross a pipeline, in addition to soil pressure it is subject to further static and dynamic loads from passing vehicles. If the soil cover is less than 1.5 m the traffic loads dominate. For cross-country pipes under railway embankments or paved roads, it is possible to calculate soil pressures and traffic loads, and hence the loads on the pipeline. The additional loads must be taken into account when determining the dimensions and properties of the pipes; special measures may be necessary.

Local conditions and climate can be particularly important for safety. In cold climates it is necessary to protect the pipeline against frost, and to elevate it on stands to prevent it from sinking and breaking during thaws. In areas subject to earthquakes the bearings must allow compensation for horizontal and vertical displacements to avoid deformation.

7.3.2. Pipeline Design

Line or network pipeline systems consist of various elements:

Pipes
Pumps
Shut-off devices (valves, stopcocks, slide gates)
Fittings
Purging cocks (for gas lines)
Condensation traps
Monitoring systems
Expansion joints

A pipeline system viewed as a whole should always be designed as a double-barrier system, capable of being monitored and repaired. The primary barrier with all its elements encapsulates the transported substance. Its freedom from leaks depends mainly on the materials used in its construction in relation to the mechanical loads to which it is subjected, and the chemical properties of the transported substances. Unavoidable losses (drip leaks at pumps, valves, and gates),

should also be contained by a second, physical barrier.

Wherever possible, there should be a continuous secondary barrier for the entire pipeline system, in the form of jacket pipes, collecting basins, and channels. In some cases this is not possible (e.g., because it would impede cathodic corrosion protection, or because differential expansion of the pipe and outer jacket threaten the system). The secondary barrier can then only consist of intensive organizational and operative back-up measures: shorter inspection periods; shorter distances between shut-off points to limit possible losses; more elaborate provisions for leak detection; continual remote monitoring of the pipeline system (both internally and externally); equipment for tackling emergency events and catastrophes; contingency plans, etc.

Such precautions are particularly important in endangered areas. Where there are threats of subsidence, it is important to carry out regular geodesic measurements along the route, and to measure the pipes for possible stretching. In densely populated areas and in important water catchment areas, the following safety measures are necessary:

– Higher specifications on the pipelines (design, materials, testing); additional supervision during pipe-laying
– Installation of additional measuring points to monitor cathodic corrosion protection
– Arrangement of shut-off fittings to limit losses
– Groundwater monitoring wells or local use of leak detectors
– Construction of safety embankments
– Laying warning bands above the overhead transmission line
– Special marking for the routes of cross-country lines
– Monitoring the route at short intervals
– Protection measures or adjacent installations

A frequent cause of incidents is damage due to external causes, e.g., excavation work. It is therefore important to provide underground pipelines with a suitable cover. For the Transalpine (TAL) Pipeline, which passes Lake Constance on its way from Genoa to Ingolstadt, the city of Stuttgart included an ozonization system in the purification plant for the water it draws from the lake, in order to be able to tackle oil leakages effectively.

7.3.3. Materials

The safety of the pipeline is affected to a considerable degree by the properties of the materials used in its construction and by quality assurance during pipe production. The measures adopted are continually being improved, e.g., better steel quality through continuous casting, secondary metallurgy, and thermomechanical rolling; improved quality control documentation (quality planning and monitoring); use of computer-aided production methods and optimum nondestructive testing procedures.

Further development of raw materials has resulted in the use of thermomechanical rolled steels and the reduction of sulfur and phosphorus levels.

When selecting and dimensioning the materials, the following criteria are important:

– Amount per unit time
– Internal pressure [constant operating pressure, shock waves (hammering), pulsating stress]
– External pressure (soil and traffic loads, bending loads)
– Temperature
– Internal and external corrosion

Selection of raw materials depends on the substances to be transported (Table 4).

7.3.4. Pipe Laying

Transport and Storage. Damage to pipes and their insulation can be avoided by suitable transport, careful loading and unloading, and proper interim storage. By training and informing personnel, careful supervision of work, and spot checks, it is possible to reduce the frequency of errors, and to detect those which do occur in time to take counter-measures.

Welding. Before welding begins in situ, it is important to check that pipes comply with delivery specifications. Only permitted filler materials may be used. Random nondestructive tests must be carried out on welded joints along the

finished pipeline section (ultrasound, irradiation, or both in combination).

Table 4. Selection of raw materials

Substance group	Special properties	Materials usually used
Water/wastewater	no special requirements (exception: corrosive water)	steel, earthenware, concrete
Gases	high mechanical strength at all pressures; resistance to corrosion by sulfur compounds and humidity in the gas	steel
Oil	high requirements for mechanical strength and expansion; high temperature stability	steel
Chemicals	high resistance to internal corrosion; high requirements for tensile strength and temperature stability	steel, copper, special alloys
Solids in suspension	high resistance to mechanical scouring; high mechanical strength	steel

Installation. In underground pipeline systems, the pipe trench and its floor must be examined before the pipe is laid. The trench must be deep enough to allow the necessary cover over the pipe, and it must be possible to install the pipe in the trench without damaging the jacket. The trench floor requires special attention to ensure that the pipe is not damaged or subjected to impermissible bending loads. The pipeline must be bedded on sand, and a compactible material must be used for filling in the trench.

7.3.5. Operation

To ensure that the pipeline transports substances at the desired rate and with constant quality, a number of measures are necessary; these depend on the necessary safety provisions, on the mode of operation, and on local conditions.

General Safety Measures. Appropriate safety equipment must be provided for all plant. The safe operation of a pipeline requires trained personnel. Organization and communication are particularly important, especially when an alarm is raised or in the event of catastrophes.

Control. The most important process information parameters, which must be continually passed to the operations center (including during interruptions), are pressure, pressure flow, volume, density, and temperature. Updates on the current state of all fittings, such as pumps, gates, relief systems, and other plant, must be passed to the operations center, together with monitoring data, in particular for the detection and location of leaks and for levels of corrosion.

In pressure pipelines, monitoring the line pressure is important. Changes in pressure can give rise to pressure waves or shock waves (hammering) which can destroy the pipeline; it is necessary to ensure continuous monitoring of shut-off valves, pressure-release devices, and safety valves.

Provisions must also be made to precisely detect, quantify, and localize losses along a pipeline due either to damage or to leaks developing under normal operating conditions. Methods are available which make use either of changes in pressure (pressure-drop or shock-wave methods) or the measurement of mass. Gradual leaks can be monitored during interruptions (pressure – temperature, pressure difference, pressure – volume methods) or during operation (leak-detection rabbits, drains, or cables).

Special rabbits or pigs can be used to identify precisely the nature, location, and size of safety threats. Distortion, corrosion, and cracks can also be identified by ultrasound, magnetic fringing flux, and high-frequency and impulse eddy current methods.

It is also important to monitor safety along the route of the pipeline, by patrols or aerial observation.

Maintenance and Repair. Regular inspection and repair of defective components serves both general safety and the operability of the pipeline systems. In addition to routine inspections on the basis of a servicing plan, a general inspection of all functions should be carried out at specific intervals (usually annually).

7.4. Incidents

No statistics on incidents are available which cover all pipeline systems and all fields of appli-

cation. In Western Europe, detailed data has been published regularly since the early 1970s by CONCAWE (oil companies' European organization for environmental and health protection) for 65 operators of cross- country oil pipelines.

There were 14 oil spillage incidents in 1991 (4 in 1990 and an annual average of 13.3 since 1971). Of these, 12 were from pipelines and 2 from pumpstations (Table 5). All incidents required cleaning up operations. The level of soil contamination caused by the incidents varied: 2 caused no contamination; 5 were classified as causing slight contamination ($< 1000\,m^2$ soil affected); 7 were classified as serious. Two petrol/diesel incidents polluted watercourses.

Table 5. Causes and numbers of incidents since 1971

	1991	Average 1971 – 91
Mechanical failure	7	3.3
Operation		0.9
Corrosion	4	4.4
Natural disaster		0.6
Damage by others	3	4.1
Total	14	13.3

The net oil loss into the environment in 1991 was $902\,m^3$ (0.00015 % of the total volume transported). The gross loss was $1346\,m^3$ (0.00023 %). A total of $444\,m^3$ (33 % of the gross loss) was recoverable.

7.5. Economic Aspects

Pipeline systems provide continuous transport and distribution of substances in bulk. Features of pipelines, in comparison with other means of transport, are low transport costs (low energy requirements), virtually loss-free transport, high operational reliability, and ease of maintenance. Pipeline systems allow greater flexibility in the choice of sites for industrial development, since the raw materials and fuels can be transported to the location, rather than the plant having to be located near the source of the raw material. Pipelines can thus be of considerable economic advantage to a country or region, since they allow the continued use of existing industrial infrastructures.

Advantages over other means of transport include shorter routes, which are usually more direct, fully automated operation (with correspondingly low costs for operation and personnel), as well as reduced environmental impact. The distances involved must be suitably large, and demand should be evenly distributed over the year, to warrant the relatively large investment costs involved.

No global summary is available for all bulk pipeline systems for oil and its derivatives, gas, water, wastewater, solids, and chemicals. In Western Europe, companies operate some 210 different service pipelines which, at the end of 1991, had a combined length of 21 000 km. In $1991, 593 \times 10^6\,m^3$ of crude oil and refined products was transported, ca. 8 % more than in 1990, most of the increase being in crude oil. This gives some idea of the importance of pipelines for the economy as a whole.

8. Future Developments

With the liberalization of shipments, world freight traffic continues to increase. In Europe this tendency is reinforced by the implementation of the Common Internal Market and by the opening up of Eastern Europe.

In the past, the transport of dangerous goods developed in parallel with freight traffic as a whole. In Germany at present about 15 % of freight volume involves the transport of dangerous goods; in the European Community the proportion is similar [103, 104]. The proportion of road transport has grown steadily, whereas rail and inland waterway transport have been pushed back (Fig. 10).

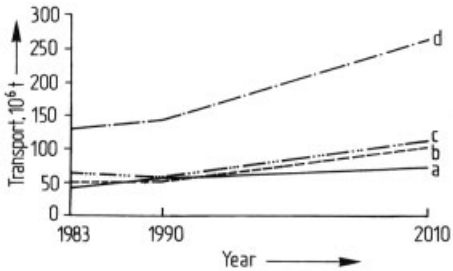

Figure 10. Transport of dangerous goods in the Federal Republic of Germany
a) Road; b) Rail; c) Ship (inland waterway); d) Total

Also to be expected in future are further increases in the transport of dangerous goods, in haulage distances, and in traffic across frontiers,

and a growth in the proportion of road transport of dangerous goods (Fig. 10).

These developments entail an inevitable increase in environmental pollution as a result of total pollutant emissions from the operation of transport systems, even if the emissions in relation to distance covered can be reduced. In addition there are waste disposal problems, such as the cleaning of freight containers. In Germany, the release of dangerous goods in traffic accidents and from storage has not increased to the same extent as the total freight traffic volume. Nevertheless, the aim must be to prevent any release of dangerous materials into the environment.

Long-distance road transport has the greatest danger potential. Figure 11, for transport of water-endangering materials in Germany, shows that long-distance road transport accounts for more than seven times as many accidents, relative to transport volume, as rail transport. The volume of water-endangering materials leaked and not recovered, relative to the transport volume, is almost three times as large. Inland waterway transport has lower accident figures than road transport, but the amount of water-endangering materials entering the environment as a result is distinctly higher than for the other carriers.

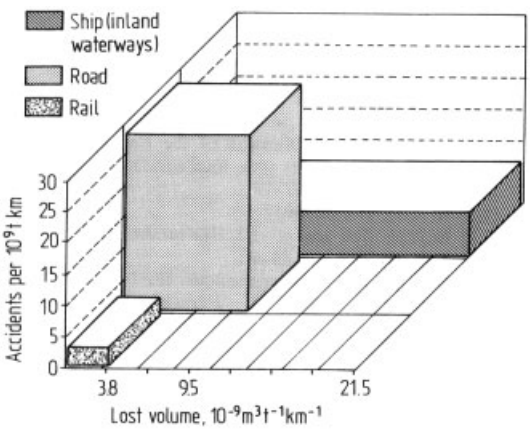

Figure 11. Potential risk of carriers for water-polluting substances

Not all shipments of dangerous goods can be transferred to rail, for capacity reasons. Besides, it may be preferable not to send a particular shipment of dangerous goods by rail if the route leads through districts that are densely populated or particularly worthy of protection, if road transport permits a harmless detour. Preconditions for a greater transfer of shipments of dangerous materials to rail are freight traffic centers on the margins of the more densely populated zones, as provided in the German Railroads plan [105], and an area-wide presence of the railroad in short-haul transport.

The existing waterways in Germany could absorb an additional freight transport volume of about 20 % [106], but the new construction or development of waterways, e.g., as investigated for the Elbe, is problematic for ecological reasons [107].

8.1. International Harmonization of Dangerous Goods Regulations

Although concern about the safety of shipments of dangerous goods has led to international agreements, questions of environmental damage have so far played rather a minor role.

On 1 January 1995 criteria for the classification of environmentally relevant substances were introduced for the first time into the international agreements on cross-border transport of hazardous materials existing since 1957 [108, 109], taking into consideration the expert recommendations of the UN.

The EU has adopted these regulations in framework directives, which remain to be introduced by the Member States. From 1 January 1997 onwards the directives will also apply to intrastate transport in the EU. Up to now the criteria deal only with aspects of water pollution; hazards to soil and air are not included.

Furthermore, there is a need for action regarding harmonization of the classification systems of the individual transport routes. Thus it is possible that goods transported by ship are classified, packaged, and labeled differently to those carried by rail or road.

In addition a uniform classification system in the field of transport should be harmonized with the classification schemes in other legal areas. Such a harmonization has already been commissioned:

In 1992 the Conference for Environment and Development (UNCED) adopted Agenda 21 in which the signatories commit themselves in

Chapter 19 to work out an internationally coordinated classification system.

8.2. Transport Avoidance and Regulatory Measures

Previous national measures, like international ones, were aimed at making the transport of dangerous goods safer: the transport volume itself was not the subject of regulation. The safest strategy for dangerous goods, however, is undoubtedly one in which they are not transported at all. Shipments of dangerous goods can become superfluous, e.g., by abandoning products involving the use of dangerous substances, or by rearranging production processes.

The best-known example of an international environmental protection agreement that resulted, as a side effect, in a rearrangement of production processes and a withdrawal of products, is the 1987 Montreal Agreement for the protection of the ozone layer. The prohibition of the production and use of specific chlorofluorocarbons that, aside from their action on the stratospheric ozone layer, are also water-endangering, has required rearrangements of production [110].

Conversion of the Basel agreement for controlling the export of wastes across frontiers to an EC statutory regulation [111] is likely to lead to an increase in the transport of wastes. This anomaly arises from the provisions in the statutory regulation, for the acceptance of new lists and reclassification and deletion of existing lists, graded according to the degree of surveillance, the vagueness of the definition of waste, and the tendency to liberalization in the worldwide movement of goods under the influence of rising costs of local waste disposal.

Aside from regulatory means, incentives for transport avoidance, and the traffic deflection, can also be achieved by economic instruments. It has been estimated that consistent application of the "polluter pays principle", which requires internationalization of external costs, would add $46\times10^9 - 250\times10^9$ DM per year to transportation costs in the federal states of pre-unification Germany. These are low estimates as they take no account of important external costs such as the effects of climatic change [112, 113].

In many cases, external costs can not be allocated unequivocally to the polluter. In addition, subsequent internalization of external costs is impossible if irreversible environmental damage has occurred [112].

The level of external costs shows that it would be possible, e.g., by a general tax on the transport of dangerous goods, to create inducements to rearrange production and contribute to transport avoidance.

A first step in this direction is represented by a proposed EC directive in which the external costs would be charged to heavy commercial vehicles (environmental loads are expressly mentioned as external costs) [114].

9. References

1. EC Green Paper on the Effect of Traffic on the Environment, Publication of the European Commission, KOM (92) 46th, final version, IV b no. 97.
2. In [1], IV a no. 87.
3. In [1], IV a no. 89.
4. In [1], III d nos. 80 – 82, III e no. 84.
5. In [1], III d nos. 83 – 85.
6. Der Rat von Sachverständigen für Umweltfragen, Umweltgutachten 1994 – Für eine dauerhaft umweltgerechte Entwicklung, Wiesbaden 1994, p. 423.
7. S. Nicodemus, H. Eberlein in Statistisches Bundesamt (ed.): *Schätzung von Umfang und Struktur des Transportaufkommens gefährlicher Güter,* im Auftrage des Bundesministers für Verkehr, Wiesbaden 1993, p. 46.
8. In [7], pp. 30 – 33.
9. The Recommendations are based on those originally prepared by the United Nations Committee of Experts on the Transport of Dangerous Goods (ST/ECA/43-E/CN.2/170) considered by the Economic and Social Council at its twenty-third session (Resolution 645 G (XXIII) of 26 April 1957). The Recommendations have been amended and updated by succeeding sessions of the Committee of Experts and published in accordance with subsequent resolutions of the Economic and Social Council.
10. Basel Convention of the Control of Transboundary Movements of Hazardous Wastes and their Disposal, 1989.
11. D. Rennoch: *Abfall-Transport-Berater,* K. O. Storck Verlag, Hamburg 1993.

12. European Agreement concerning the International Carriage of Dangerous Goods by Road (ADR) and protocol of signature done at Geneva on 30 September 1957, United Nations Publication, ECE/TRANS/110, vols. I and II New York – Geneva.Ordnung für die internationale Eisenbahnbeförderung gefährlicher Güter (RID), Zentralamt für den internationalen Eisenbahnverkehr, Bern.
13. International Maritime Dangerous Goods Code (IMDG-Code), International Maritime Organization, London 1995.
14. *Laboratory Biosafety Manual,* WHO, Geneva 1983.
15. OECD Guidelines for Testing of Chemicals, no. 404, Acute Dermal Irritation/Corrosion, Paris 1992.
16. *Official Journal of the European Communities* **34** (1991) 180.
17. American Society of Mechanical Engineers: *Boiler and Pressure Vessel Code,* Section VIII, Division 1, "Rules for the Construction of Pressure Vessels," ASME, New York 1992.
18. *AD-Merkblätter,* Beuth Verlag GmbH, Berlin 1994.
19. B. Jäger et al.: *Die Auswirkungen des Sicherheitsrisikos von Lagerung und Transport gefährlicher Stoffe auf die Entwicklung verbesserter Transporttechnologie,* BMFT-Report RGB 8010, Verlag TÜV Rheinland 1983.
20. Health and Safety Commission: *Major Hazard Aspects of the Transport of Dangerous Substances,* Library and Information Services, London 1991.
21. M. Arras, *TU* **34** (1993) 84.
22. E. Lehmann et al., *TU* **34** (1993) 79.
23. Underwriters Laboratories Inc.: Outline of Proposed Investigation of Structural Steel Protected for Resistance to Rapid-Temperature-Rise-Fires, UL 1709, London, Sep. 1983.
24. B. Droste, *Fire Technol.* **28** (1992) 257 – 269.
25. *Allianz-Handbuch der Schadenverhütung,* VDI-Verlag, Düsseldorf 1984.
26. *Handbook of Industrial Loss Prevention,* McGraw-Hill Book Co., New York 1967.
27. W. E. Baker et al.: Workbook for Estimating Effects of Accidental Explosions in Propellant Ground Handling and Transport Systems, NASA Contractor Report 3023, National Aeronautics and Space Administration, 1978.
28. Control of Urban Development Around High-Risk Industrial Sites, Secretary of State to the Prime Minister for the Environment and the Prevention of Major Technological and Natural Risks – DEPPR, France, Oct. 1990.
29. G. Uelpenich, *Brandschutz* **47** (1993) 696 – 702.
30. Verein Deutscher Ingenieure (ed.): VDI 3783, part 1, Dispersion of Emissions by Accidental Releases – Safety Study, 1987; part 2, Dispersion of Heavy Gas Emissions by Accidental Releases – Safety Study, 1990, Beuth Verlag, Berlin 1987, 1990.
31. Verband der Chemischen Industrie e.V. (VCI), TUIS, Transport-Unfall-Informations- und Hilfeleistungs-System, Frankfurt/Main, Oct. 1992.
32. CEFIC, ICE News, Brussels, Belgium, Sep. 1992 and May 1993.
33. H. Wefers et al.: Hinweise zur Erstellung und Prü-fung von betrieblichen Alarm- und Gefahrenabwehrplänen nach der Störfall-Verordnung, LIS-Berichte no. 83, Essen 1988.
34. Recommendations on the Transport of Dangerous Goods, United Nations Publication, New York.
35. Recommendations on the Transport of Dangerous Goods, Tests and Criteria, United Nations Publications, New York.
36. Regulations for the Safe Transport of Radioactive Material, International Atomic Energy Agency, Vienna.
37. International Maritime Dangerous Goods Code, Volumes I to IV and Supplement, International Maritime Organization, London.
38. Technical Instructions for the Safe Transport of Dangerous Goods by Air, International Civil Aviation Organization, Montreal.
39. Règlement concernant le transport international ferroviaire des marchandises dangereuses (RID), Office Central des Transports Internationaux par Chemins de Fer, Bern.
40. European Agreement concerning the international carriage of dangerous goods by road (ADR), United Nations, Publication, ECE/TRANS/100, New York.
41. Recommendations on the Transport of Dangerous Goods, ST/SG/AC. 101/Rev 8, latest edition United Nations Publication, New York.
42. International Maritime Dangerous Goods Code (IMDG-Code), 27th Amendment, International Maritime Organization, London 1995.

43. European Agreement Concerning the International Carriage of Dangerous Goods by Road (ADR), United Nations Publication, ECE/TRANS/100, vols. I and II, New York – Geneva, 1994.
44. European Agreement Concerning the International Carriage of Dangerous Goods by Rail (RID), unofficial edition by Department of Transport, HMSO Publications Centre, London 1995.
45. Technical Instructions for the Safe Transport of Dangerous Goods, International Civil Aviation Organization, DOC 9284-AN/905, latest edition, Montreal.
46. Regulations for the Safe Transport of Radioactive Materials, 1985 Edition (as amended 1990), Safety Series no. 6, STI/PUB/517, International Atomic Energy Agency, Wien.
47. Advisory Material for the IAEA Transport Regulations for the Safe Transport of Radioactive Material, 1985 Edition, Safety Series no. 37, STI/PUB/589, International Atomic Energy Agency, Wien 1990.
48. Europäisches Übereinkommen vom 20. Sept. 1957 über die internationale Beförderung gefährlicher Güter auf der Straße (ADR) in der Fassung der Bekanntmachung vom 15. Januar 1992 (BGBl: II, S. 95) in der Fassung der 11. Änderungsverordnung vom 4. März 1993, BGBl. II S. 234 – Anlagenband zum BGBl. Teil II Nr. 9 vom 23. März 1993.
49. Ordnung für die internationale Eisenbahnbeförderung gefährlicher Güter (RID), Anlage I zu Anhang B des Übereinkommens über den internationalen Eisenbahnverkehr vom 9. Mai 1980 (COTIF-Übereinkommen) zuletzt geändert durch die 4. RID-Änderungsverordnung vom 25. 06. 1993 (BGBl. II Nr. 19).
50. International Maritime Dangerous Goods Code (IMDG-Code), 25th Amendment, International Maritime Organization, London 1989; B. Anz. Nr. 69, 8 April 1992.
51. Verordnung über die innerstaatliche und grenz-überschreitende Beförderung gefährlicher Güter auf Straßen (GGVS); Europäisches Übereinkommen über die internationale Beförderung gefährlicher Güter auf der Straße (ADR).
52. Verordnung über die innerstaatliche und grenzüberschreitende Beförderung gefährlicher Güter mit Eisenbahnen (GGVE); Ordnung über die internationale Eisenbahnbeförderung gefährlicher Güter (RID).
53. H. Engels: "Tankcontainer minimieren Risiko," *Chem. Ind.* (1988) no. 10, 28 – 32.
54. E. Mühling: "Safety First," *Chem. Ind.* (1990) no. 1, 12 – 14.
55. B. Schulz-Forberg, B. Droste: "Tankbau – was, wie und warum," *Gefährliche Ladung* **36** (1991) no. 11, 506 – 512.
56. K. Wissenbach: "Gefährliche Güter sicher transportieren," *Maschinenschaden* **62** (1989) no. 5, 181 – 185.
57. Recommendations on the Transport of Dangerous Goods, 5th ed., United Nations Publication, New York 1988.
58. K. Wissenbach: "Gefährliche Güter sicher transportieren," *Maschinenschaden* **62** (1989) no. 5, 181 – 185.
59. M. Hegenauer: "Vorschriften für die Beförderung gefährlicher Güter," *Chem.-Tech.* **14** (1985) no. 6, 153 – 164.
60. K. Ridder: "Transportvorschriften für gefährliche Güter," *Entsorgungs-Technik,* Jan./Feb. 1990, 36 –38, March/April 1990, 38 – 40
61. V. Pilz: "Technische Maßnahmen zur Verhinderung von Schadstoff-Freisetzung aus Chemie-Anlagen und beim Chemikalien-Transport," *VFDB Z.* **35** (1986) no. 4, 160 – 163.
62. H. J. Mayer: "Für und Wider von Überdruckventilen an Tanks zur Beförderung gefährlicher Güter," *TU* **22** (1981) no. 6, 245 – 248.
63. R. J. Janssen: "Machen Überdruckventile Druckgaskesselwagen sicherer?" *Gefährliche Ladung* **25** (1980) no. 9, 376 – 378.
64. G. F. List: "Modeling and Analysis for Hazardous Materials Transportation," *Transp. Sci.* **25** (1991) no. 2, 100 – 114.
65. International Maritime Organization (IMO), SOLAS-Convention 74/78 as amended.
66. IMO, MARPOL-Convention 73/78 as amended.
67. Zentralkommission für die Rheinschiffahrt (ZKR), ADNR Regulations – Version prior to 1995.
68. ZKR, ADNR Regulations – Version 1995.
69. Germanischer Lloyd (ed.): "Inland Waterway Vessels," part 2 in "I Ship Technology," *Rules for Classification and Construction.*
70. Germanischer Lloyd (ed.): "Seagoing Ships," part 1 in "I Ship Technology," *Rules for Classification and Construction,* Chap. 1.

71. Germanischer Lloyd (ed.): "Seagoing Ships," part 1 in "I Ship Technology," *Rules for Classification and Construction,* Chaps. 6, 7.
72. K. J. Döhrn: "Das neue ADNR/ADN, Bauvorschriften für Binnentankschiffe," Germanischer Lloyd, Hamburg.
73. M. Böckenhauer: "Die aktuellen MARPOL-Vorschriften zur Verbesserung der Tanksicherheit," Germanischer Lloyd, Hamburg..
74. VDI-Richtlinie 2700, "Ladungssicherung auf Straße," Mai 1990.
75. G. Großmann et al.: *Technologie für Transport, Umschlag und Lagerung im Betrieb,* VEB Verlag Technik, Berlin 1983.
76. N. Müller: "Neuland für Stückgut-Fahrer," *Gefährliche Ladung* **2** (1993) 66.
77. ISO 1496 – 4:1991, Series 1 freight containers – specification and testing – Part 4: Non-pressurized containers for dry bulk.
78. DIN 30 723,Absetzkipperfahrzeuge, Juli 1986.
79. DIN 30 722,Abrollkipperfahrzeuge, Abrollbehälter, November 1990.
80. Richtlinien zur Durchführung der Gefahrgutverordnung Straße vom 24. März 1994, VkBl., no. 7, 1994, p. 281.
81. Richtlinien für die sichere Behandlung von Schüttladungen bei der Beförderung mit Seeschiffen vom 30. August 1990, B Anz., Nr. 226 a vom 6. Dez. 1990, zuletzt geändert durch die im B Anz., Nr. 23 vom 04. Feb. 1993 veröffentl. Bekanntmachung vom 18. Dez. 1992.
82. P. Schmitt: *Lagerung gefährlicher Güter in der Europäischen Gemeinschaft,* Verlag TÜV Rheinland, Köln 1991.
83. P. Roberts: "Application of Hazard and Risk Analysis to Gas Making and Gas Storage Facilities," International Conference on Safety and Loss Prevention in the Chemical and Oil Processing Industries, 1990.
84. J. E. S. Venart: "Major Hazards in the Transport and Storage of Pressure Liquefied Gases," *J. Hazard. Mater.* **12** (1988) 3 – 392.
85. V. Pilz: "Sichere Lagerung von Stoffen in der chemischen Industrie – safe storage of chemicals," *Chem. Ing. Tech.* **60** (1988) no. 6, 452 – 463.
86. V. Marshall: *Major Chemical Hazards,* Ellis Horwood Ltd., Chichester 1987.
87. V. Pilz: "Technische Maßnahmen zur Verhinderung von Schadstoff-Freisetzungen aus Chemie-Anlagen und bei Chemikalien-Transporten," *VFDB Z.* **35** (1986) no. 4, 160 – 163.
88. K. J. Doktor, A. Borchard: "Abnahmeuntersuchung an Druckbehälteranlagen," *TU* **30** (1989) nos. 7/8, 294 – 296.
89. K. J. Doktor: "Gase sicher lagern," *TU* **30** (1989) no. 6, 245 – 249.
90. D. L. Blomquist: "Neue Normen für Schadenverhütung bei Flüssiggasen (API 2510 and 2510 A)," *Hydrocarbon Process.* **11** (1988) 58 – 61.
91. API Standard 2510 – Design and Construction of Liquefied Petroleum Gas (LPG) Installations, 6th ed., American Petroleum Institute, Washington, D.C. 1989.
92. API Standard 2510 A – Fire Protection Considerations for the Design and Operation of Liquefied Petroleum Gas (LPG) Storage Facilities, 1st ed., American Petroleum Institute, Washington, D.C. 1989.
93. K. Strohmeier: "Abschätzung des Gefährdungspotentials druckverflüssigter Gase," *Chem. Ing. Tech.* **65** (1993) no. 4, 410 – 413.
94. A. Schwierczinski, R. Herrmann: "Praxisbezogene Ermittlung von Sicherheitsabständen bei der Ausbreitung freigesetzter schwerer Gase," Siemens AG (KWU), Offenbach 1990.
95. "Ermittlung sicherheitstechnischer Kriterien zur Flüssiggastechnologie und Herleitung geeigneter Sicherheitsstandards," Bundesanstalt für Materialforschung und -prüfung (BAM), Berlin 1988, Report 01501, VDI-Technologiezentrum Physikalische Technologien.
96. DIN 4119,Oberirdische zylindrische Flachboden-Tankbauwerke aus metallischen Werkstoffen, Teil 1 Grundlagen, Ausführung, Prüfungen, Teil 2 Berechnung, Beuth Verlag GmbH, Berlin 1979/1980.
97. Verordnung über brennbare Flüssigkeiten – VbF mit ihren Anhängen und den dazugehörigen Technischen Regeln für brennbare Flüssigkeiten (TRbF), Carl Heymanns Verlag KG, Berlin 1992.
98. BS 7777 – Flat-bottomed, Vertical, Cylindrical Storage Tanks for Low Temperature Service; part 1, Guide to the general provision applying for design construction, installation and operations; part 2, Specification for the design and construction of single, double and full containment metal tanks for the storage of liquified gas at temperatures down to −165 °C; part 3, Recommendations for the design and

construction of prestressed and reinforced concrete tanks and tanks foundations; and the design and installation of tank insulation, tank liners and tank coatings; part 4, Specification for the design and construction of single containment tanks for the storage of liquid oxygen, liquid nitrogen and liquid argon, British Standard Institution, London 1992.

99. Technische Regeln zur Druckbehälterverordnung (TRB), Carl Heymanns Verlag KG, Köln, loose leafe collection.
100. O. Scharz: *Glasfaserverstärkte Kunststoffe, kurz und bündig,* Vogel-Verlag, Würzburg 1975.
101. BS 4994, Design and Construction of Vessels and Tanks in Reinforced Plastics, British Standard Institution, London 1987.
102. K. Krekeler, G. Wick: "Polyvinylchlorid," *Kunststoffhandbuch,* vol. I. Carl Hanser Verlag, München 1993.
103. S. Nicodemus: *Transport von gefährlichen Gütern 1982 – 1991,* Statistisches Bundesamt, Wiesbaden 1992.
104. A. Coco: "Grand Prix d'Harmonisation," *Gefährliche Ladung* **7** (1991) 278.
105. Die Deutschen Bahnen, Masterplan GVZ Deutschland, Zentrale der DR/DB, Berlin/Frankfurt a.M. 1993.
106. G. Hulsmann: "Die Binnenschiffahrt in großer Not," *Frankfurter Allgemeine Zeitung,* June 18, 1993.
107. Öko-Projekt Elbe-Raum, Elbe – Fluß oder Kanal, Tagungsband, Dresden 1992.
108. ADR, Accord européen relativ au transport international des marchandises Dangereuses par voie de Route; BGBl, 1969 II, p. 1489.
109. (RID), Office Central des Transports Internationaux par Chemins de Fer, Bern; BGBl, 1985 II, p. 666. Réglement International concernant le transport feroviaire des marchandises dangereuses.
110. Umweltbundesamt: *Verzicht aus Verantwortung: Maßnahmen zur Rettung der Ozonschicht,* Berichte 7, Erich-Schmidt-Verlag, Berlin 1989.
111. Beschluß des Rates vom 1. Febr. 1993 zum Abschluß des Übereinkommens über die Kontrolle der grenzüberschreitenden Verbringung von gefährlichen Abfällen und ihrer Entsorgung (Basler Übereinkommen), 93/98/EWG.
112. E. Dogs, H. Platz: *Externe Kosten des Verkehrs – Schiene, Straße, Binnenschiffahrt,* Planco Consulting GmbH, Essen 1990.
113. D. Teufel et al.: *Ökologische und soziale Kosten der Umweltbelastung in der Bundesrepublik Deutschland im Jahr 1989,* Umwelt- und Prognose-Institut, Heidelberg 1991.
114. Änderung des Vorschlags für eine Richtlinie des Rates zur Anlastung der Wegekosten an schwere Nutzfahrzeuge, Drucksache der Europäischen Kommission, KOM (92) 405 endg., vom 30.09.1992.

Waste

Karl J. Thomé-Kozmiensky, Institut für Technischen Umweltschutz, Technische Universität Berlin, Berlin, Federal Republic of Germany (Chaps. 1, 3, Sections 9.3, 9.4)

Sophie Willnow, Berlin, Federal Republic of Germany (Chap. 2)

Günter Fleischer, Institut für Technischen Umweltschutz, Technische Universität Berlin, Berlin, Federal Republic of Germany (Chap. 4 in part, Chap. 6, Sections 7.1, 7.5)

Wulf-Peter Schmidt, Institut für Technischen Umweltschutz, Technische Universität Berlin, Berlin, Federal Republic of Germany (Chap. 4 in part)

Claus Christ, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany (Chap. 5)

Georg Menges, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Federal Republic of Germany (Section 7.2)

Bernd Bilitewski, Ingenieurgesellschaft für Technischen Umweltschutz, Dresden, Federal Republic of Germany (Sections 7.3, 7.4)

Ulrich Loll, Ingenieurbüro für Abwasser-, Abfall- und Aquatechnik, Darmstadt, Federal Republic of Germany (Section 7.6)

Hubert Gromotka, Hessische Industriemüll GmbH, Wiesbaden, Federal Republic of Germany (Chap. 8)

Norbert Amsoneit, Zweckverband Sondermüll-Entsorgung Mittelfranken, Schwabach, Federal Republic of Germany (Section 9.1)

Manfred Baerns, Institut für Angewandte Chemie, Berlin-Adlershof e. V., Berlin, Federal Republic of Germany (Section 9.2)

Manfred Baerns, Lehrstuhl für Technische Chemie, Ruhr-Universität Bochum, Bochum, Federal Republic of Germany (Section 9.2)

Frank Majunke, W. C. Heraeus GmbH, Hanau, Federal Republic of Germany (Section 9.2)

Hans-Jürgen Ehrig, Bergische Universität Gesamthochschule Wuppertal, Wuppertal, Federal Republic of Germany (Section 10.1)

Hans-Joachim Schneider, BRP Consult, Berlin, Federal Republic of Germany (Section 10.2)

Volkmar Gossow, Bilfinger + Berger Umweltbau GmbH, Köln, Federal Republic of Germany (Section 10.3)

1. Origin of Waste 6
1.1. Municipal Waste 8
1.2. Hazardous Waste 14
1.3. Future Determination of Waste Generation 16
2. Legal Aspects 17
2.1. Introduction 17
2.2. The Recycling Act 18
2.3. Principles of Recycling 19
2.3.1. Avoidance before Recovery 19
2.3.2. Material and Thermal Reclamation 20
2.4. Legal Formulation of Recycling 20
2.5. Disposal Standard – TA Abfall . 21
2.5.1. TA Abfall 21
2.5.2. TA Siedlungsabfall 22
2.6. Hierarchy of Responsibility . . . 23
2.7. Approval of Waste Disposal Facilities 23
2.8. Monitoring of Waste Disposal . 24
2.8.1. Monitoring 24
2.8.2. Transportation Approval 25
2.9. Plant Officer for Waste 26
2.10. Waste Management Concepts and Balances 26
2.10.1. Waste Management Concepts . . 26
2.10.2. Waste Management Balances . . . 26
2.11. Remediation of Hazardous Waste Sites 26
2.11.1. Duty to Remediate 26
2.11.2. Other Duties 27

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b08_559

2.11.3. Financing of Remediation of Hazardous Waste Sites 27
2.12. Criminal Offenses and Infringements of Regulations 28
2.13. Civil Liability 28
2.13.1. Product Responsibility 28
2.13.2. Damage Claims 29
3. Organization of Hazardous-Waste Disposal 29
4. Life-Cycle Assessment 32
4.1. Introduction 32
4.2. Scope and Limits of Life-Cycle Assessments 33
4.3. Method of Product Life-Cycle Assessment 34
4.3.1. Goal Definition and Scope 34
4.3.1.1. Knowledge Interest and System Description 34
4.3.1.2. Functional Unit 35
4.3.2. Inventory Analysis 35
4.3.2.1. Components of the Product Line 36
4.3.2.2. Development of the Product Line 36
4.3.2.3. Data Acquisition 36
4.3.2.3.1. Process Description 38
4.3.2.3.2. Acquisition of Basic Data 38
4.3.2.4. Material and Energy Balance . . . 38
4.3.2.4.1. Process Balance 38
4.3.2.4.2. Process Modeling and Data Correction 39
4.3.2.5. Allocations 40
4.3.2.5.1. Process Product and Coproducts . 40
4.3.2.5.2. Object of Investigation 41
4.3.2.5.3. Open-Loop Recycling 41
4.3.2.5.4. Equality of Benefits 42
4.3.3. Impact Assessment 43
4.3.3.1. Classification 43
4.3.3.2. Characterization 43
4.3.3.2.1. Greenhouse Effect 44
4.3.3.2.2. Ozone Depletion 45
4.3.3.2.3. Acidification 45
4.3.3.2.4. Eutrophication 45
4.3.4. Interpretation 46
4.3.4.1. Objective and Requirements . . . 46
4.3.4.2. Interpretation Approaches 46
4.3.4.2.1. Normalizing, Quantifying Methods 46
4.3.4.2.2. Verbal-Argumentative Assessment 47
4.3.4.2.3. Reductionistic Approach 48
4.3.5. Improvement Assessment (Optional) 48
4.4. Critical Review 48
5. Waste Management in the Chemical Industry 49
5.1. Introduction 49
5.2. Chemical Industry Wastes . . . 49
5.3. Waste-Management Concepts . 50
5.3.1. Residues and Wastes from Production 50
5.3.2. Production-Oriented Management 50
5.3.3. Product-Oriented Management . 51
5.4. Disposal Measures 52
5.4.1. Logistics 52
5.4.2. Waste Combustion 52
5.4.3. Landfill Disposal of Wastes 53
5.4.4. Asbestos Disposal 55
5.5. Utilization of Product Wastes . 55
5.5.1. Plastics Recycling 55
5.5.2. Refrigerant Recycling 56
5.5.3. Recycling of Used Packaging Materials 58
5.5.4. Paint Recycling 58
5.6. Results of Waste Management . 58
6. Prevention of Waste 60
6.1. Introduction 60
6.2. Distinction between Primary and Secondary Waste Prevention 62
6.3. Waste Prevention and Reduction of Environmental Burdens 63
6.4. Specific Actions to Reduce Environmental Burdens 66
6.4.1. Decreasing Resource Consumption 66
6.4.2. Reduction of Waste and Emission Quantities 68
6.4.3. Reduction of Pollution Potential . 68
6.4.4. Extending Product Service Life . 68
6.4.5. Improvement of Recyclability . . 69
6.5. Integrated Measures to Reduce Environmental Burdens 70
6.5.1. Product-Integrated Environmental Protection 70
6.5.2. Production-Integrated Environmental Protection 71
6.6. Determination of Environmental Burdens 72
7. Recycling of Waste 73
7.1. Introduction 73
7.1.1. General Aspects 73
7.1.2. Forms of Recycling 73
7.1.3. Prerequisites for and Limitations of Recycling 76
7.1.4. Selection of Disposal Route . . . 78
7.1.5. Recycling Rates 78
7.2. Plastics 79
7.2.1. Introduction 79

7.2.2. Mechanical Recycling 80
7.2.2.1. Chemical Basis for Plastics Recycling 80
7.2.2.2. Pure Wastes 81
7.2.2.3. Mixed and Soiled Plastics Wastes 82
7.2.2.4. Fiber Composites and Other Thermosets 83
7.2.2.5. Outlook for Mixed Wastes in Blends 83
7.2.3. Degradation to Molecular Constituents 83
7.2.3.1. Biodegradable Plastics 83
7.2.3.2. Chemical Recycling 84
7.2.3.2.1. Chemical Degradation 84
7.2.3.2.2. Thermal Degradation 85
7.2.4. Combustion (Energy or Thermal Recycling) 91
7.2.4.1. Combustion in Waste Incinerators 92
7.2.4.2. Shredded Waste as Fuel for Sewage Sludge 92
7.2.4.3. Combustion with Pure Oxygen . 93
7.3. Solvents 93
7.3.1. Introduction 93
7.3.2. Processing of Solvent Wastes . . 94
7.3.2.1. General Aspects 94
7.3.2.2. Mechanical Separation of Solids 95
7.3.2.3. Treatment of Halogenated Solvents 95
7.3.2.4. Treatment of Halogen-Free Solvents 96
7.4. Paints 99
7.4.1. Introduction 99
7.4.2. Recycling of Paint Residues . . . 100
7.4.2.1. General Aspects 100
7.4.2.2. Internal Paint Recycling 100
7.4.2.3. External Paint Recycling 101
7.5. Metals 103
7.5.1. Introduction 103
7.5.2. Production Waste Recycling . . . 105
7.5.3. Recycling during Product Use and Product Reconditioning 108
7.5.4. Product Waste Recycling 108
7.5.4.1. Product-Specific Recycling 109
7.5.4.1.1. Automobiles 109
7.5.4.1.2. Catalytic Converters 112
7.5.4.1.3. Lead – Acid Batteries 113
7.5.4.1.4. Dry-Cell Batteries 115
7.5.4.1.5. Electronics Scrap; Circuit Boards 116
7.5.4.2. Material-Specific Recycling . . . 118
7.5.4.2.1. Steel 119
7.5.4.2.2. Aluminum 119
7.5.4.3. Recycling of Material Wastes . . 120
7.5.5. Recycling of Residues 123
7.6. Sludges 124
7.6.1. General 124
7.6.2. Quantities of Sewage Sludge Treated 124
7.6.3. Requirements for Sludge Recycling and Disposal 126
7.6.4. Sewage Sludge Treatment Processes 126
7.6.5. Unit Operations in Sewage Sludge Treatment 126
7.6.5.1. Thickening and Dewatering 126
7.6.5.2. Stabilization 128
7.6.5.3. Disinfection 130
7.6.5.4. Conditioning 130
7.6.5.5. Drying 131
7.6.6. Quantities of Other Sludges;Disposal Options 132
7.6.6.1. Industrial Production Sludges . . 132
7.6.6.2. Water Treatment Plant Sludges . 132
7.6.6.3. Dredging Sludges 133
7.6.6.4. Sludges from Mining 134
8. Intermediate Storage of Hazardous Wastes 134
8.1. Introduction 134
8.2. Intermediate Storage Sites . . . 135
8.3. Zone Structure of an Intermediate Storage Site 135
8.4. Procedures at an Intermediate Storage Site 137
8.5. Soil Protection 138
8.6. Conclusions 139
9. Treatment of Waste 139
9.1. Physicochemical Treatment . . . 139
9.1.1. Introduction 139
9.1.2. Types, Origin, and Properties of Wastes 140
9.1.3. Treatment Processes 140
9.1.4. Requirements for Construction and Operation of Physicochemical Treatment Plants 141
9.1.5. Technological Concept of Physicochemical Treatment Plants . . . 142
9.1.5.1. Treatment Plant for Inorganically Contaminated Wastes 142
9.1.5.2. Treatment Plant for Organically Contaminated Wastes 143
9.1.6. Environmental Requirements . . . 144
9.1.7. Storage and Transportation 146
9.1.8. Legal and Economic Aspects . . . 147
9.2. Hydrolysis and Alcoholysis of Plastics Wastes 147
9.2.1. Introduction 147
9.2.2. State of the Art 147
9.2.2.1. Hydrolysis of Polyurethanes . . . 148
9.2.2.2. Alcoholysis of Polyurethanes . . 151

9.2.2.3. Hydrolysis and Alcoholysis of Polyesters and Polyamides 152
9.2.2.4. Alcoholysis of Mixed Plastics . . 154
9.2.3. Conclusions and Outlook 154
9.3. Incineration 154
9.3.1. Grate Firing 157
9.3.2. Fluidized-Bed Firing 162
9.3.3. Rotary-Kiln Firing 167
9.4. Pyrolysis and Gasification 172
9.4.1. Principles of Waste Pyrolysis . . . 172
9.4.2. Principles of Gasification of Waste 177
9.4.3. Processes 179
9.4.3.1. Pyrolysis – Combustion Process . 179
9.4.3.2. Thermoselect Process 183
9.4.3.3. Noell Conversion Process 191
10. Deposition 198
10.1. Deposition Aboveground 198
10.1.1. Introduction 198
10.1.2. Classification of Landfills 198
10.1.3. Landfill Siting 199
10.1.4. Landfill Safety Analyses 199
10.1.5. Lining and Drainage Systems . . 200
10.1.6. Leachate and Gas Formation in Landfills 201
10.1.6.1. Processes in the Landfill 201
10.1.6.2. Water Balance 203
10.1.6.3. Leachate Quality 203
10.1.6.4. Leachate Treatment 205
10.1.6.5. Gas Formation 207
10.1.6.6. Landfill Gas Recovery 208
10.1.7. Landfill Operation 208
10.2. Underground Deposition 208
10.2.1. Introduction 208
10.2.2. Types of Mines and Their Relevance to Waste Deposition and Recycling 210
10.2.3. Safety Requirements for Operation and Closure of an Underground Deposition Site or a Mine Subject to Stowage 211
10.2.4. Wastes and Residues for Deposition in Mined Cavities 212
10.2.5. Examples of Spaces Used for Waste Deposition and Spaces Packed with Residues Not Generated in Mining 213
10.3. Sealing of Existing Sites 214
11. References 218

The entire topic was coordinated by Karl J. Thomé-Kozmiensky

1. Origin of Waste

In Germany quantities of waste generated must be determined according to the German Environmental Statistics Act. The Federal Statistical Service (StBA) [1] publishes two summaries:

1) Statistics on public waste disposal
2) Statistics on waste disposal in production plants and hospitals

The two figures are based on different principles. In the StBA waste balance, waste disposal data are compared with waste production data. Duplicate entries are removed, so that a net balance is presented without substances passed on to reprocessing businesses and the scrap trade.

Wastes produced in 1990 amounted to some 300×10^6 t, which broke down as follows:

Sewage sludge	2 %
Hazardous waste	3 %
Household waste	15 %
Industrial waste	39 %
Construction waste	41 %

Construction wastes are the major component. The net contribution of hazardous wastes, according to the waste balance, is ca. 10×10^6 t/a and the gross amount ca. 15×10^6 t/a.

The StBA statistics have been criticized on the following grounds [2]:

1) The results appear too late
2) The figures are not adequately broken down
3) The data are not in accord with present-day constraints on waste management
4) Waste-prevention and reclamation activities are not covered by the statistics
5) The StBA waste balances are incomplete

The same problems exist throughout the field of waste statistics.

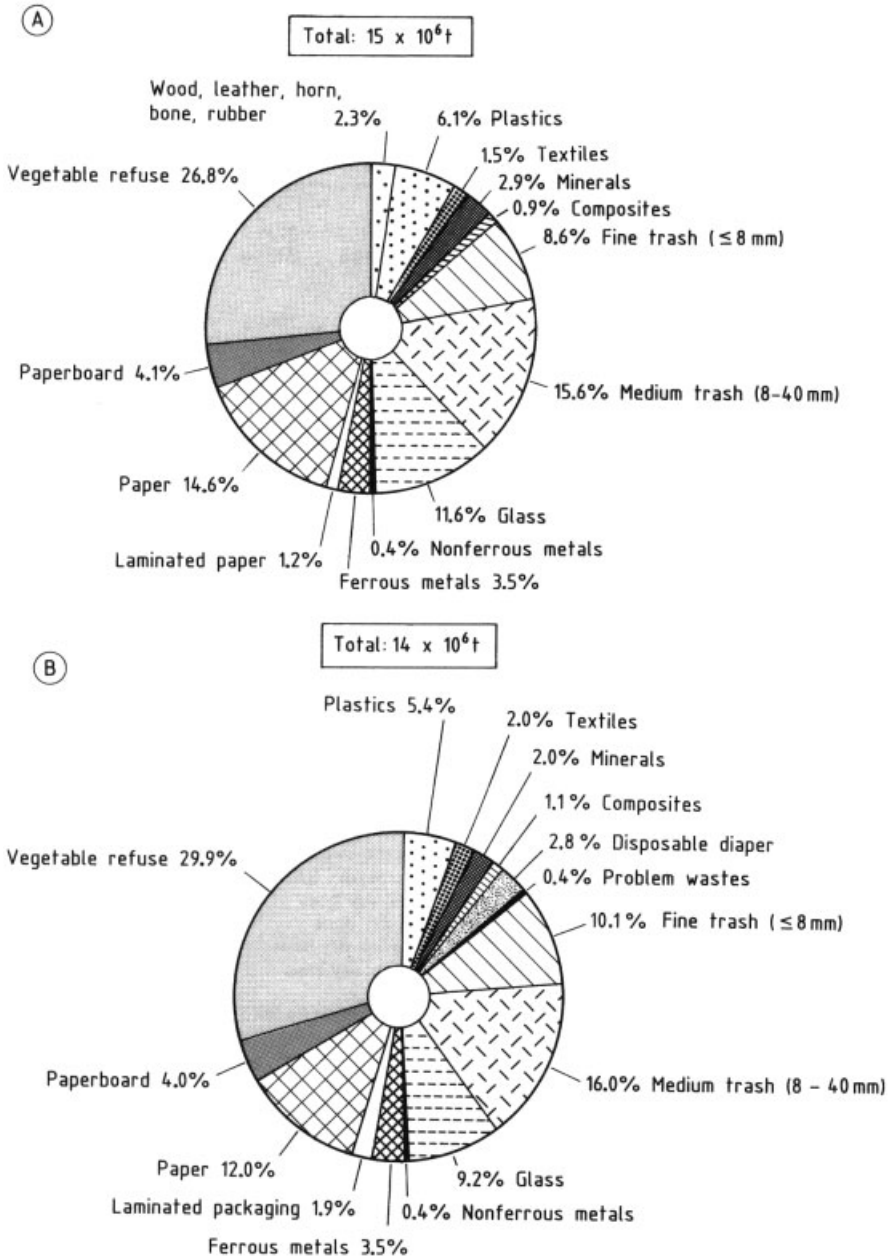

Figure 1. Composition of residential waste in 1979 – 1980 (A) and 1985 (B) (in wt %) [4, p. 17]

1.1. Municipal Waste

Table 1 summarizes the quantities of household, commercial, institutional, and bulky wastes collected in public waste-disposal activities [3].

Household waste includes all wastes and residues from chiefly private households, which are regularly collected, transported, and disposed of in the interest of keeping dwellings clean.

Table 1. Quantities of residential, commercial, institutional, and bulky waste colleted in public waste-disposal activities [3]

Land	Total quantity collected		Commercial waste separately collected/removed ∗		Commercial waste not separately collected (per capita) ∗∗		Amount (10^3 t) collected by	
	1000 t	1000 m^3	1000 t	1000 m^3	kg	L	Public agencies	Private companies
Pre-unification Federal Republic								
1984	22 118	137 454			362.3	2 251.7	11 553	10 565
1987	22 895	148 215			374.6	2 425.0	11 696	11 199
Germany								
1990 ∗∗∗	30 461	183 561	3 941	17 466	333.2	2 086.8	13 673	16 788
Baden-Württemberg	2 617	17 455			266.4	1 777.1	724	1 893
Bavaria	3 126	20 622	102	417	264.2	1 764.9	1 190	1 936
Berlin	1 546	8 408	388	1 640	337.4	1 970.9	1 258	288
Brandenburg	1 303	5 099	84	233	478.6	1 911.0	684	618
Bremen	364	2 348	24	109	499.0	3 282.9	189	176
Hamburg	924	7 981	264	1 797	399.1	3 742.5	660	264
Hesse	2 247	12 786	318	1 571	334.8	1 946.0	1 105	1 143
Mecklenburg-West Pomerania	944	5 649	27	165	478.0	2 860.4	13	931
Lower Saxony	2 722	19 409	148	838	348.5	2 514.0	1 494	1 228
North Rhine-Westphalia	6 502	43 486	383	2 177	352.7	2 381.0	4 150	2 353
Rhineland-Palatinate	1 705	7 208	400	1 551	346.8	1 503.0	594	1 112
Saarland	426	2 580	31	202	367.5	2 216.6	178	248
Saxony	2 828	12 062	1 186	4 262	347.7	1 651.1	642	2 186
Saxony-Anhalt	1 031	5 879	99	372	333.8	1 971.5	–	1 031
Schleswig-Holstein	1 094	6 344	151	776	359.2	2 120.3	456	638
Thuringia	1 081	6 246	337	1 356	285.0	1 873.4	338	743
Former Federal Republic	22 759	143 475	2 208	11 077	322.5	2 077.6	11 480	11 279
Former Democratic Republic	7 702	40 086	1 733	6 388	376.2	2 123.8	2 193	5 509

∗ Information not collected to 1987.
∗∗ Including separate commercial collection prior to 1987.
∗∗∗ Preliminary result.

Table 2. Per capita waste generation and gross national product , pp. 25 – 26[1]

	Municipal waste, kg/a		Gross national product, $	
	1986	1990	1985	1990
United States	744	864	16 700	21 607
Finland		608		
Denmark	399	469		
Holland	500	467	8 460	18 708
Switzerland	336	427	14 940	33 941
Japan		394	11 030	23 973
Great Britain	335	353	8 060	16 912
Germany	374	331	10 230	24 070
Spain	275	322		
Sweden	301	317	11 900	27 281
Greece		314		
Belgium		313	8 090	19 528
France	260	304	9 300	21 055
Italy	249	301	6 270	18 729
Portugal	167	231		
Austria	216	228	8 670	20 906

This includes:

1) Food wastes and kitchen wastes if not carried away with sewage
2) Waste paper
3) Packaging materials from consumer goods
4) Heating residues (e.g., ashes)
5) Small objects that have become unusable
6) Broken crockery, glass, porcelain, etc.

When countries are ranked in terms of waste generated per inhabitant, Germany falls in the middle part of the list (Table 2). In interpreting these figures, however, note that usage of the word “waste” differs from country to country.

Figure 1 ([4, p. 17]) shows the makeup of residential waste based on the German federal household waste analysis in 1979 – 1980 and 1985. Some characteristics of household waste are listed in the following [1]:

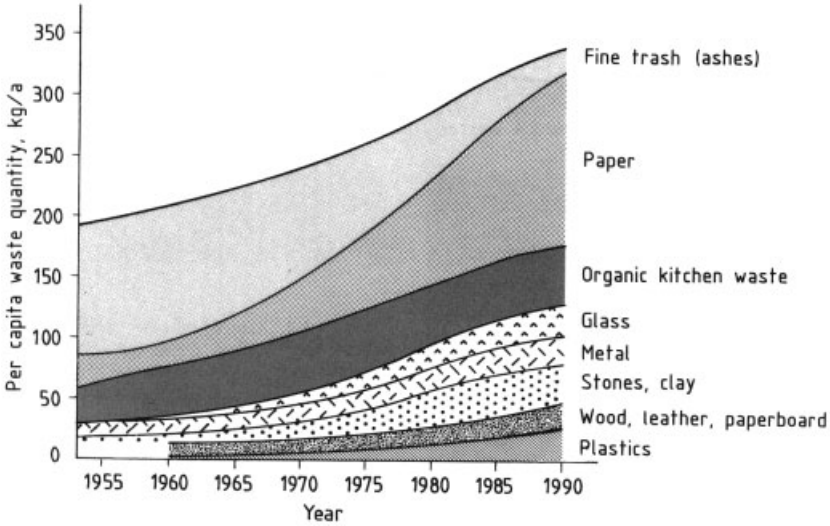

Figure 2. Waste composition over time [4, p. 18]

Lower heating value, kJ/kg	6000 – 10 000
Combustile content, wt %	45
Ash (inerts), wt %	25
Water, wt %	30
Merchant iron (rel. to total wt.), wt %	4 – 5
Bulk density, kg/m^3: loose	90 – 120
compacted (landfill vehicle)	350 – 550
Organic matter as TOC, wt %, dry basis	30 – 70
Inorganic matter, wt %, dry basis	30 – 70
Carbon, wt %, dry basis	10 – 50
Hydrogen, wt %, dry basis	4 – 5
Nitrogen, wt %, dry basis	1 – 2
Chlorine, wt %, dry basis	0.1 – 1.0
Sulfur, wt %, dry basis	0.03 – 0.5
Lead, g/kg, dry basis	0.7 – 2.0
Cadmium, g/kg, dry basis	0.01 – 0.05
Mercury, g/kg, dry basis	0.001 – 0.05

Figure 2 shows how the composition of residential waste has changed over the past decades.

Specific waste volume (i.e., volume per inhabitant) has increased much more rapidly than specific waste weight since 1950 (Fig. 3). From about 1900 until after World War II, the amount of waste generated – both the weight and the volume per inhabitant and year – remained almost constant. Marked decreases were recorded during the two World Wars. With the coming of prosperity starting around 1950, the Länder of the pre-unification Federal Republic showed a continuous increase in quantity of waste produced, whereby the specific weight increased less rapidly than the specific volume. The reasons for this discrepancy were the shift from fires in single-room stoves to central heating and the increased use of packaging, plastics, and printed materials. In the former German Democratic Republic, the weight and volume curves ran nearly parallel, even after World War II. The situation in the new eastern Länder changed after reunification. The per capita amount of waste produced was for a time much greater than in the western Länder (pent-up demand), but now seems to be swinging back into line with the western Länder. The use of combustible wastes in stoves and furnaces is declining steadily as oil- and gas-fired central and district heating systems are established, while the quantities of packaging and printed paper waste are similar.

As to the quantities of valuable products identified in German residential waste and recovered by sorting, figures for individual substance classes are rough, and the balances do not work out exactly for all substances.

The quantities available and recovery rates of Table 3 were estimated on the basis of the German household waste analysis of 1985, which found 14×10^6 t of residential waste to be disposed of, together with a scenario and an adjustment for 1993 packaging statistics [6].

Table 3. Forecast recovery of values from household waste [a] in Germany, 1993 [6]

Substance class	Household waste, t	Quantity disposed of, t	Quantity recovered, t	Recovery rate [d], %
Paper, paperboard [b]	7 500 000	3 700 000	3 800 000	50.6
Glass	3 700 000	1 300 000	2 400 000	63.2
Aluminum	130 000	120 000	10 000	8.0
Tinplate	700 000	450 000	250 000	35.7
Textiles	600 000	325 000	275 000	45.8
Plastics	1 400 000	1 100 000	300 000	21.4
Vegetable material	11 800 000	10 300 000	1 600 000	13.6
Residual waste [c]	3 100 000	3 100 000		
Total	29 030 000	20 395 000	8 635 000	29.8

[a] Waste from private households, excluding bulky and commercial waste. [b] Packaging paper and graphic paper. [c] Fine trash mineral material, mineral composites, etc. [d] Not identical to rates of German Packaging Regulation.

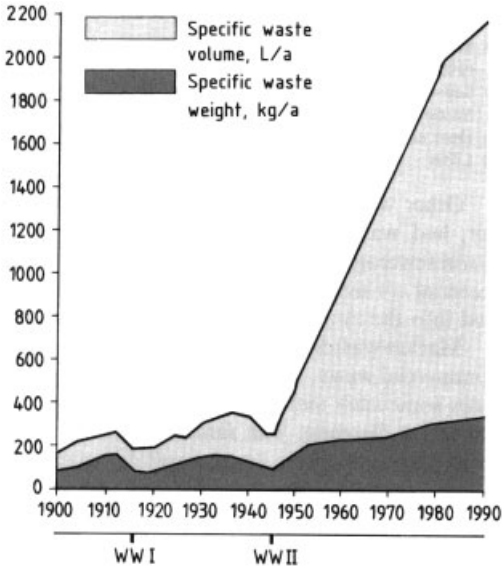

Figure 3. Specific waste volume and weight over time (1900 – 1945 for Germany, later for pre-unification Federal Republic) [4, p. 18]

Commercial waste has increased faster than household waste (see Fig. 4). Biowaste includes organic, compostable constituents of residential waste that are collected separately and composted. Problem wastes (pollutants) contained in residential waste are small quantities of substances that cause harm to humans, the environment, and facilities. Such substances include solvents, paints, batteries, medications, pesticides, and household chemicals.

Bulky waste from households belongs to household waste and is related to it, differing only in size. The boundary between residential and bulky waste is fluid, depending chiefly on the size of the collection containers. Bulky waste, even after size reduction, generally cannot be collected along with the ordinary containers because of its bulky nature.

According to BILITEWSKI, the factors governing production of bulky waste include the following [5]:

1) Size of the municipality (area and population)
2) Fraction of commercial waste collected together with residential waste
3) Consumption and lifestyle patterns, standard of living, economic status
4) Commercial, social and geographic structure
5) System used for residential waste collection: container size and type (container volume per inhabitant and day)
6) Number of persons in the building from which waste is collected
7) Frequency and organization of waste collection
8) Waste producers' alternatives to ordinary waste disposal: special containers for valuable materials, self-composting, recycling yards, recycling pickups by associations, collection by private contractors
9) Fees charged for residential and bulky waste collection

Garden and Park Waste. Garden waste is generated in gardens, parks, and cemeteries and also on streets (e.g., leaves, branches, grass, and clippings) and is commonly not removed by waste collection. It poses a problem for treatment facilities handling household waste, because it arises periodically and in large quan-

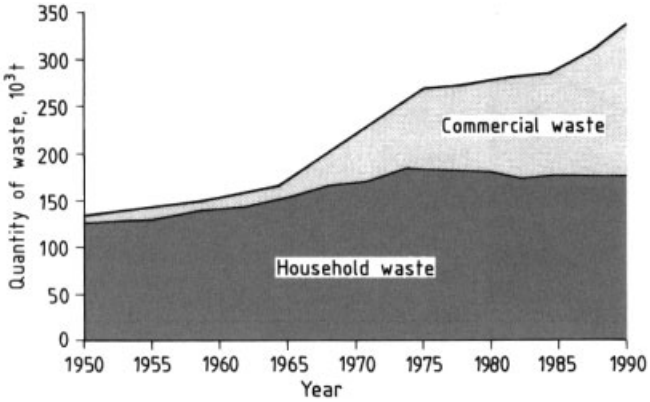

Figure 4. Quantity of waste produced by a large city, broken down into household/bulky waste and commercial waste [4, p. 18],

tities. Where garden waste cannot be recycled by composting on the spot, it is usually removed on an individual basis.

Market Waste. Street waste consists of

1) Material washed and rubbed off street surfaces plus rubber from vehicle tires
2) Leaves and branches dropped by streetside trees
3) Material scattered for winter traction
4) Dust emitted by private and commercial furnaces
5) Litter

Other wastes include trash, cans, waste paper, leaf waste, weeds, and solids retained in storm sewerage intakes (catch basin sludge). This material consists chiefly of street sweepings carried into the intakes by rain or sprayed water.

Market wastes are included in the category of commercial waste, but because of special market rules some such material is collected by municipal street cleaning. The same applies to waste from public markets (e.g., vegetable and leaf waste, packaging, and shipping materials such as wooden crates and pallets).

Construction Waste. Rubble and excavated soil are commercial wastes that are not generally collected together with other wastes, but they may represent a large fraction of the total waste produced in a region (depending on construction activitiy).

Rubble includes broken masonry, mortar residues, and concrete, and often contains metal fragments, wood, and paper waste as well. This category also includes furnace lining material, slag, sand from metal foundries, and sludges from the production of concrete, cement, mortar, and limestone sand. These wastes can be processed and recycled.

Excavated soil comprises natural soil and rock, not detrimentally modified, that is dug out or scraped off during excavation and earthmoving. If such soil is expected to contain pollutants, it must be analyzed and treated separately. Excavated soil can be used as fill or to cover other wastes in landfills.

Mixed construction site wastes are mixtures of rubble, concrete material, and excavated soil of mineral origin, as well as paper, plastics, wood, glass, ceramics, insulants, metals, and pollutants. They are generated by new construction, repair work, and renovation.

The quantities of various construction wastes are listed in Tables 4 and 5.

Table 4. Generation of construction waste in the Federal Republic of Germany , p. 45[5]

Type of waste	Amount generated ∗, 10^6 t/a	Recovery rate ∗∗, wt %
Excavated soil	170	32 (∗∗∗)
Building rubble	23	16 (60)
Pavement rubble	20	55 (90)
Construction site waste	12	0 (40)

∗ Extrapolated to 1990.

∗∗ Figures in parentheses are targets set by the German Federal Government.

∗∗∗ Uncontaminated soil should be reclaimed and should not be dumped in landfills.

Table 5. Forecast generation of construction waste in the former Democratic Republic, 1990 , p. 45[5]

Type of waste	Amount generated, 10^6 t/a	Estimated recovery rate prior to 1990, wt %
Excavated soil	5	
Pavement rubble	6 – 7	
Building rubble	5 – 6	2
Construction site waste		
Total	16 – 18	

Table 6. Composition of Berlin sewage sludge (dry basis) , p. I/52[7]

Parameter	Unit	Fresh sludge	Digested sludge
Loss on ignition	wt %	75	62.7
C	wt %	40	34
H	wt %	5.4	4.8
O	wt %	24	18.2
N	wt %	4	3.5
S	wt %	1.6	2
Cl	wt %	0.3	0.3
F	mg/kg	1	1
As	mg/kg	2.6 –	25.0
Cd	mg/kg	1.6 –	21.0
Co	mg/kg	3.2 –	5.0
Cr	mg/kg	19.6 –	1200.0
Cu	mg/kg	87.9 –	860.0
Hg	mg/kg	1.6 –	4.7
Mn	mg/kg	74.0 –	260.0
Ni	mg/kg	94.0 –	340.0
Pb	mg/kg	52.0 –	124.0
Sb	mg/kg	1.8 –	10.0
Sn	mg/kg	54.0 –	673.0
Tl	mg/kg	0.6 –	19.0
V	mg/kg	9.5 –	25.5
Zn	mg/kg	657.0 –	2500.0
$\sum$ AOX	mg/kg	192.0 –	287.0
$\sum$ PCB	µg/kg	80.0 –	846.0
PCDD/F	TE ng/kg	15.0 –	470.0

Sewage Sludge. Household sewage sludge consists of a mixture of water with solids separated from wastewater in communal clarifying and treatment plants.

Because communal sewage treatment plants also handle industrial wastewaters, household sludge in the narrow sense is only produced in domestic sewage treatment systems and consists chiefly of human excrement.

The sludge retained in mechanical wastewater clarifiers is referred to as clarifier sludge. The sludge from trickling filters is called trickling filter sludge; that from activated-sludge processes is called activated sludge.

Sludge from a modern municipal treatment plant with a biological stage consists of a mixture of clarifier and activated sludges.

There are two forms of sludge:

1) Fresh residential sewage sludge (fresh or raw sludge) is obtained as top or bottom sludge from clarifiers.
2) Digested residential sewage sludge (digested sludge) is sludge that has been subjected to an anaerobic biological process in special digesters.

An example of sewage sludge composition is given in Table 6. The content of pollutants must be considered in assessing possible applications of the sludge; its water content and heating value has to be taken into account with regard to self-sustaining combustion.

Commercial and industrial wastewaters are generated directly or indirectly in connection with the commercial objective of a company or the production of industrial goods.

These wastewaters are classified according to business size:

1) Wastewaters from small commercial operations are often released into municipal sewerage without pretreatment. They are diluted by residential wastewater in the sewers. They usually have little effect on the composition of municipal wastewater.
2) Wastewaters from larger commercial and industrial sites can crucially alter the quantity and quality of municipal wastewater. They must be allowed for in the design of municipal sewage treatment plants. These wastewaters are often pretreated in on-site clarifiers before being released into municipal sewerage.

Commercial and industrial sludges are residues from clarifying systems in industrial plants. Their composition depends on the production process as well as the treatment. They vary wide in nature, including:

1) Lime mud from neutralizers
2) Hydroxide sludge from chemical precipitation systems
3) Hair and skin scraps from tanneries
4) Sand and soil from washing of potatoes, carrots, etc.

Solids from industrial wastewaters are separated along with residential sewage sludge in municipal treatment plants. Because of their special physical or toxic properties, they interfere with utilization of the sludge. Sludges containing metal salts, cyanides, and so forth are banned for agricultural use, for example. Excess sludge from biological treatment systems in industry contain primarily dissolved organic contaminants; as far as composition is concerned, these can be equated to activated sludge from household wastewater.

1.2. Hazardous Waste

The amount of hazardous wastes produced can be determined in two ways:

1) Through triennial surveys conducted by the Federal Statistical Service on waste disposal in producing industries
2) From waste waybills used as an aid in monitoring waste transportation

Neither procedure can dependably supply exact data. Sources of error include the following:

1) When waste waybills are tabulated, duplications are often missed, because the records show hazardous wastes en route from the producer to physicochemical treatment and also individual constituents and components en route from pretreatment to landfill
2) The Federal Statistical Service surveys only cover wastes produced by operations with more than 19 employees

Waybill information was tabulated for the Federal Republic of Germany in 1984, 1985, and 1988. The quantity of hazardous waste increased in the period from 1984 to 1988. The total in 1988 was 10.2×10^6 t, including 3.3×10^6 t of hazardous wastes requiring nationwide certification [2].

Figure 5 shows quantities of hazardous waste based on the 1984, 1985, and 1988 tabulations and 1990 figures of the StBA.

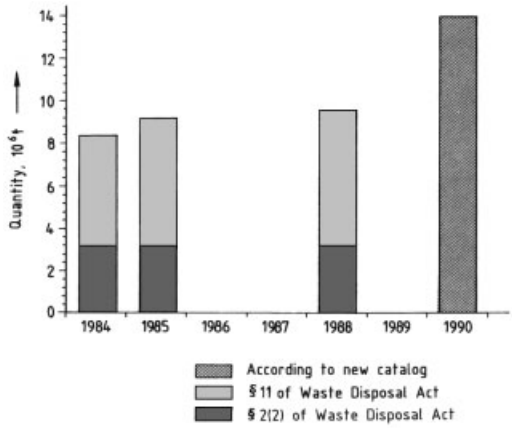

Figure 5. Quantities of hazardous waste based on tabulation of waybills

Additional hazardous waste is generated by pollutants present in residential waste. Some 700 tons of hazardous wastes was collected separately in 1993.

According to the German Technical Directive on the Disposal of Hazardous Waste (TA Sonderabfall), the following treatment routes are available for hazardous wastes:

1) Physicochemical and biological treatment
2) Incineration
3) Disposal in aboveground landfills
4) Underground disposal

Annex C of the TA Abfall gives a catalog of wastes particularly subject to regulation. The only wastes listed are those from the annex of the German Waste Definition Regulation (Abfallbestimmungs-Verordnung). The catalog gives instructions on physicochemical, biological, or thermal treatment as well as disposal of various types of waste. Also included are defective intermediate and end products that may be generated by production processes.

Residual materials listed in the annex of the German Residual Material Definition Reg-

ulation (Reststoffbestimmungs-Verordnung) are not included in this catalog. A type of waste that is essentially recoverable will appear in the catalog of substances particularly subject to regulation only if a lack of reclamation options dictates that the waste has to be disposed of.

The disposal instructions apply only in cases where the reclamation test of TA Abfall, Section 1.3, gives a negative result and the waste must be disposed of pursuant to Section 1.4. The catalog of wastes particularly subject to regulation is organized along mixed lines on the basis of the properties, composition, and state of aggregation, as well as the origin of the waste class.

In this catalog, high-level sorting concepts are used to categorize waste classes in three successive, systematic levels called supergroup, group and subgroup. The decimal classification is used to identify these categories:

Supergroups	one-digit numbers
Groups	two-digit numbers
Subgroups	three-digit numbers

The five-digit numbers (waste codes) contained in the subgroups characterize the classes of waste. Every waste subject to disposal must be assigned one of these waste codes.
Example:

1	Supergroup: Wastes of plant and animal origin, wastes of refined products
17	Group: Wood wastes
17211	Subgroup: Sawdust and chips, oil-impregnated or with harmful contaminants, chiefly organic

The catalog is based on nomenclature customary in waste management. The names of supergroups, groups, and subgroups are part of the names of waste classes. They must be taken into account when determining the class of waste. When pure substance names are selected to denote a waste class, the pure substance is the waste-defining component even though the waste generally contains impurities.

In other cases, environmentally relevant properties are included in the name of a waste class, for example physical and chemical properties.
Example:

17211	Sawdust and chips, oil-impregnated or with harmful contaminants, chiefly organic
17212	Sawdust and chips, with harmful contaminants, chiefly inorganic

Some waste class names include terms such as "free of" or "without." These terms are not absolute but relate to the environmental relevance of the admixtures. Information as to origin is intended as exemplary and provides additional description of the waste class.
Example:

57801	Shredder residues (light fraction)

The regulatory relevance depends on the quantity. The waste class can be found in the catalog through a summary of supergroups, groups, and subgroups, or through an alphabetical index. If a waste can be assigned distinct waste codes, the preferred key is the one that better characterizes the waste with regard to its properties.

Columns 5 to 11 of the catalog contain disposal instructions based on experience:

CPB:	Physicochemical/biological treatment
HMV:	Municipal waste combustor
SAV:	Hazardous-waste incinerator
HMD:	Municipal waste landfill
SAD:	Hazardous-waste landfill
UTD:	Underground deposition
Others:	Special handling (for example)

When there are several disposal options, preference classes are entered in columns 5 – 10. Preference class 1 identifies the normal disposal case; preference class 2 identifies a disposal route that is also possible because of:

1) The composition of the waste
2) Special measures, for example special pretreatment of the waste or extra precautions in the disposal facility
3) Site- and plant-specific circumstances

When the disposal options rank equally, the same class number is repeated.

For waste classes where the pollutant levels can fluctuate widely, the reference under preference class 2 is to options associated with residential waste: HMV or HMD. This is the case for wastes that can be disposed of in an environmentally compatible fashion in waste-disposal facilities not covered by TA Abfall.

1.3. Future Determination of Waste Generation

Quantities of wastes produced will be determined more exactly in the future. Advances are expected through improved surveys when the new German Environmental Statistics Act comes into force and through balancing by substance from winning of raw materials through waste production.

New Environmental Statistics Act. With the promulgation of the Environmental Statistics Act (Umweltstatistikgesetz) on September 28, 1994, years of preparation and discussion of practicable waste statistics come (provisionally) to an end. The act embodies compromises between opposing viewpoints in some instances.

The material part of the new act governs surveys and data acquisition in four areas:

1) Waste management
2) Water resources
3) Air pollution abatement
4) Environmental economics

The act establishes a number of broad rules for meeting the requirements of the Federal Statistics Act (Bundesstatistikgesetz).

In deference to the federal consolidation program, the act does not take effect until January 1, 1997; in the meantime, the provisions of the 1980 Environmental Statistics Act remain in force.

The following goals are the new act:

1) Data acquisition is adapted to an increased demand for information.
2) Information will be kept more up-to-date.
3) Reporting obligations are unified so that preparation of reports will be less onerous.

The priority task now is methodological and organizational preparation for surveys.

Substance-by-Substance Balancing. If exact information about waste production is to be obtained in the future, this is possible only if individual substances are balanced from production to reclamation. This requirement is also in accord with the view of the Council of Specialists on Environmental Questions [8]. WUTTKE has presented an example of batchwise balancing for PCBs [2]. The PCB stream for 1992 is given as follows:

Accumulated domestic production in Germany was 76 000 t; imports amounted to 27 000 t. The total domestic consumption of PCBs is thus 103 000 t. Exports were 16 000 t, leaving 87 000 t in the country. This quantity was divided between open and closed application routes.

Open application in printing inks, plastics, self-copying papers, etc. accounted for ca. 25 000 t of PCBs. Open application has severely contaminated all environmental media. For example, PCBs in self-copying paper passes via residential waste into waste treatment; with old pa-per, through paper recycling, into new paper and into the wastewater of plants processing old paper. PCBs were also used in joint-sealing compounds for concrete buildings and in flame-retardant coatings for ceiling tiles, and in this way later becomes a contaminant in building rubble.

Closed application, for example in electrical capacitors and transformers and in hydraulic fluids accounted for ca. 62 000 t of PCBs. This application has led, and continues to lead, to environmental and workplace burdens all along the reclamation route, as is shown by the presence of PCBs in lubricating oil systems and in steel scrap.

Because of the ban on PCBs, more PCB-containing devices and materials are now being disposed of. But not all the PCBs produced and used are disposed of. PCBs from most open applications have already passed into the environment in dispersed form, though in part they have already been biologically degraded. Also, PCBs reach disposal facilities not just in their own form but also in devices and materials, so that the quantity of PCB waste yet to be disposed of is still thought to be large.

After 1993 and up to the year 2000 (corresponding to the term of the German Chemical Ban Regulation), the Federal Environmental Office expects ca. 340 000 t of PCB waste to be disposed of. Possible disposal routes are incineration for liquid wastes (over 130 000 t) and underground deposition for solid wastes (ca. 105 000 t).

The discrepancies result from uncertainties in the data, as would be expected given the na-

ture, quantity, and duration of PCB use, which was formerly not subject to reporting. According to the Federal Environmental Office, 20 000 t of PCB-containing wastes from the rehabilitation of buildings with concrete joint sealants containing PCBs must be disposed of. These materials, however, will not be disposed of until the year 2000; the anticipated amount for the next six years is 4000 – 5000 t.

2. Legal Aspects

2.1. Introduction

The German Recycling Act became law in the version of Bundestag Document 12/8084 and was promulgated in the autumn of 1994. It is a so-called article act. Article 1 contains the law as such, and is titled "Gesetz zur Förderung der Kreislaufwirtschaft und Sicherung der umweltverträglichen Beseitigung von Abfällen (Kreislaufwirtschafts- und Abfallgesetz – KrW/AbfG)" [Act to Promote Recycling and to Ensure the Environmentally Safe Disposal of Wastes]. Article 13 governs the coming into force of all provisions. As a rule, this occurs two years after promulgation of the act, that is, in autumn 1996. Until then, the Waste Act (Abfallgesetz) of 1986 remains in force.

As soon as the Waste Act – "Gesetz über die Vermeidung und Entsorgung von Abfällen (Abfallgesetz – AbfG)" [Act on the Avoidance and Disposal of Wastes] – took effect in 1986, the scene was set for the implementation of a waste disposal concept attuned to the times.

While it had been common to collect all wastes and dispose of them in landfills, possibly after pretreatment, the 1986 Waste Act set new priorities: avoidance of waste takes priority over waste recycling and disposal.

This objective corresponds to the principles of a modern waste management policy: The generation of wastes must be avoided in the industrial sector if possible. Any wastes produced should be recycled. Wastes that cannot be avoided or recycled must be disposed of without harm to the environment.

The Waste Act of 1986 has been expanded in significant ways, such as the authorization to issue general administrative regulations, the new provisions for the disposal of used oil, the requirement to keep certain harmful substances separate and the seller's obligation to take them back, and the possibility of setting goals for reducing the problem of waste quantities. These changes have been set forth in a series of regulations, administrative directives, and target. Important for the chemical industry in particular are the Waste Determination Regulation [Abfallbestimmungs-Verordnung], the Residual Substance Determination Regulation [Reststoffbestimmungs-Verordnung], and the Technical Directive on Wastes [TA Abfall].

In addition, it has been or still is necessary to take account of several international codes such as the EC Directives in drafting bills and regulations.

Nevertheless, waste management is confronted with problems not yet solved: In spite of all measures to avoid and recover wastes, quantities of waste continue to increase. Exporting larger amounts of waste is not a solution, since there is no guarantee for their environmentally safe disposal. The disposal of hazardous wastes poses special difficulties. Low public acceptance of new disposal facilities of this type has led to a shortage of disposal capacity.

These above aspects, the associated rises in disposal costs, and difficulties in obtaining permits under the Federal Pollution Control Act – which may be issued only for disposal of hazardous wastes under safeguards – are making hazardous waste disposal an even more serious problem for companies. At first, hazardous waste landfills had to accept all problem wastes without safeguards. Today, these landfills (such as Gerolsheim in Rhineland-Palatinate, Georgswerder in Hamburg, and Münchehagen in Lower Saxony) have become some of the most dangerous polluted sites, and so it is certain that hazardous waste disposal must be decisively improved.

This chapter therefore sets forth and discusses all principles of law relevant to the chemical industry. Waste disposal is carried out subject to the German Federal Waste Act and supplementary state laws. The regime of waste law and its fundamental concepts must therefore be explained first. The most important problem areas, such as disposal, its organization, and its financing, are then discussed against the background of the pertinent legislation.

2.2. The Recycling Act

According to the concurrent German Federal Legislation as anchored in the German Constitution, the Federal States have the right to pass their own waste disposal legislation only as long as and to the extent that the Federal Republic has not made use of its right to legislate. In the time following passage of the Federal Waste Disposal Act (Abfallbeseitigungsgesetz) in 1972, all Federal States adopted state waste management laws, sometimes amending their own prior legislation; these acts are substantially alike in that they are more or less implementing acts to the Federal Waste Act.

The Recycling Act can be applied only when the property in question is waste in the sense of the act. The concept of "waste" is therefore crucial. The meaning of this term in German Federal Law (§ 3 Sect. 1 Clause 1 KrW/AbfG) represents a conclusive rule that state and local legislation can neither narrow nor broaden. The new definition of waste, in comparison with the definition used in the 1986 Waste Act, implies a substantially broader potential applicability of waste law. In principle, all substances and objects that are not products are subject to surveillance pursuant to waste law (§ 40).

> According to § 3, waste comprises all movable property
> – That falls under the classes enumerated in Annex 1 (objective component) and
> – Whose owner is getting rid of, wishes to get rid of, or must get rid of it (subjective component)

This concept of waste corresponds to that of EC Council Directive 91/156, though the directive speaks not of "movable property" but of "substances or objects."

The definition makes it clear, that waste can only be movable property. This immediately excludes property that is not movable, such as real estate, contaminated soil before its excavation, and houses before their demolition.

The term "owner" used in the definition is legally defined by Art. 1 Letter c of EC Directive 75/242 as amended. Beyond mere possession, it also takes in the intention of exercising such possession. The term "owner" is also legally defined in § 3 Sect. 6. The owner of wastes is any natural person or corporate body having actual possession of wastes.

The definition of waste in European law, like the German definition, includes the constituent fact "get rid of" or "riddance." This term is defined by § 3 Sect. 2. This text distinguishes three cases of riddance:

1) Reclamation in the sense of Annex II B
2) Disposal in the sense of Annex II A
3) The surrender of actual possession of movable property with no further determination of purpose

"Delivery" in the sense of the first two cases means the direct handing over of the property to an entity prepared to reclaim or dispose of it. The surrender of actual possession with no further determination of purpose means the handing over of the property to persons and institutions that in turn perform treatment operations necessary for recovery or disposal, such as haulers. The third case of riddance provides that anyone who surrenders actual possession of movable property with a definite end in view, for example, sells or transports it, does not "get rid of" the property in the sense of § 3 Sect. 2.

The concept of waste has an objective component: it covers waste in the objective sense. Wastes in this sense are all movable property that fall into the classes enumerated in Annex I and whose owner must get rid of them. The waste classes correspond to the waste classes of EEC Directive 91/156. The Commission has issued the European waste catalog, filling these classes, by a decree of December 20, 1993. As a result of the enumeration principle, there can be no problems with regard to movable property falling under recycling law. The same holds for the obligation of riddance: it is defined in legislation. According to § 3 Sect. 4, the obligation of riddance exists when the movable property is no longer being used according to its originally intended purpose; if its concrete condition is such that it is suitable to endanger the welfare of the public, in particular the environment, in the present or future; and if its hazard potential can be eliminated only through an orderly and innocuous reclamation or disposal compatible with the public welfare pursuant to the provisions of this act and of regulations issued on the basis of this act. The criterion "obligation of riddance" of the objective concept of waste con-

tains normative elements. The movable property can (1) no longer be used in accordance with its original purpose – this is the case if the user can no longer employ the property according to its intended purpose, that is, objectively the property has lost its originally intended purpose. The concrete condition of the property (2) must be such that it is suitable to endanger the welfare of the public, in particular the environment, in the present or future – this suitability is to be established in accordance with objective criteria. The hazard potential (3) can be eliminated only through an orderly and innocuous reclamation or disposal compatible with the public welfare pursuant to the provisions of recycling law; this question is also answered on the basis of objective criteria.

The concept of waste has a also subjective component: it covers waste in the subjective sense. Only the property listed in the waste catalog can become waste in the subjective sense. In this way, however, no gap is created: The waste catalog contains a "cushion" criterion. The closed character of the catalog therefore does not imply the exclusion of certain property from the waste concept. As before, anything can become waste in the subjective sense. The objects or substances of the waste catalog are waste in the subjective sense if the owner gets rid of them – decisive for this is the above-stated behavior in the sense of § 3 Sect. 2 – or wishes to get rid of them. The intention to get rid of property is defined in § 3 Sect. 3. The standard distinguishes two cases: Intention to get rid of property (1) is presumed for such movable property that arises in energy conversion, manufacturing, processing, or utilization of substances or articles or in services, without the purpose of each operation being directed hereto. For property not comprised by the purpose of the operation, the intention of riddance is thus constructive; economically exploitable movable property that arises in the activities named in the act is waste in the subjective sense. The intention of riddance (2) is presumed when the original intended purpose of movable property falls away or is abandoned without a new purpose of use arising directly at its location. One can say that an intention of riddance is to be presumed if the movable property has become meaningless for the owner. In order to assess the intended purpose according to § 3 Sect. 3 Clause 1 No. 2, the view of the producer or owner is to be considered, with allowance for the normal view, according to § 3 Sect. 3 Clause 2. According to the definition of § 2 Sect. 5, "producer" means any natural person or corporate body that has performed prior operations that effect a change in the nature or the composition of these wastes. What is waste in the subjective sense, according to § 3 Section 3 Clause 2, is substantially governed by the subjective view of the producer or owner. In contrast to previous waste law, there is no separate document defining hazardous waste or wastes subject to special surveillance. According to the definition in § 3 Sect. 1, these wastes are also covered by the reference to the waste classes enumerated in Annex I.

2.3. Principles of Recycling

The first principle is that of domestic disposal. All wastes in the sense of the act, according to § 10 Sect. 3, must be disposed of within the country. Pertinent EC provisions, however, remain unaffected (Regulation on Supervision and Control of Shipments of Wastes within, into, and out of the EU; Act Implementing the Basel Convention of March 22, 1989, on the Monitoring of Cross-Border Transportation of Hazardous Wastes and Their Disposal).

According to § 4 Sect. 1, wastes are first to be avoided and only second is their material or energetic content to be recovered. This statement is in accord with the corresponding provision of the EC Framework Directive on Waste.

2.3.1. Avoidance before Recovery

The objective is made clear in § 4 Sect. 2 by means of a description of waste avoidance practices. These include closed-cycle operation within plants, low-waste product design, and appropriate consumption behavior. This listing makes it clear that the attainment of the goal of waste avoidance depends on the understanding and fulfillment of product responsibility.

With regard to production, the Recycling Act clearly distinguishes between the plant-oriented Federal Pollution Control Act and more strongly substance-oriented waste law, in that § 9 indicates that plant-oriented requirements may be

implemented only according to the provisions of the Pollution Control Act. This holds, in particular, also for internal plant operation with substances in closed cycles. The Pollution Control Act also presumes that plants produce wastes that must be either avoided or else reclaimed in an orderly and innocuous manner.

2.3.2. Material and Thermal Reclamation

The delimitation of material and thermal reclamation is essentially based on § 4 Sect. 1 No. 1, which sets a clear priority for avoidance before any type of reclamation. Only when avoidance is not possible, in particular because of closed-cycle operation within a plant or the reduction of the quantity and pollutant level of waste, the second requirement for reclamation comes into play. A legal definition of material reclamation appears in § 4 Sect. 3; of thermal reclamation, in § 4 Sect. 4.

Material Reclamation. Three forms of material reclamation are distinguished:

1) The replacement of raw materials by recovering substances from wastes (legal definition: "secondary raw materials"). Example: manufacture of pulp from recycled paper to cut down new pulp production (conservation of forests).
2) Utilization of the material properties of wastes for the original purpose. Example: production of motor oil from waste oil.
3) Utilization of the material property of wastes for other purposes with the exception of direct energy extraction. Example: use of composted sewage sludge in agriculture.

The boundaries between modes of material reclamation are often fluid. Even an analysis of whether the principal purpose of the measure lies in the utilization of waste or the elimination of its pollutant potential (§ 4 Sect. 3 Clause 2) often can be decided only on a case-by-case basis. Abstractly, the only further prerequisite for reclamation is that the waste has been treated at all. Accordingly, reclamation is only the economically significant use of wastes for new application areas after physical, chemical, or biological treatment. The concretization of the material reclamation modes and their priority with respect to thermal reclamation, to be assessed under global ecological criteria, will have to be carried out in the framework of regulations oriented toward substances or industrial sectors.

Thermal Reclamation. According to § 4 Sect. 4, energetic reclamation includes the use of wastes as a substitute fuel. The preference for energetic reclamation over disposal means that thermal treatment of wastes in order to dispose of them, particularly of household waste, is ruled out. The distinction between energetic reclamation and thermal treatment is handled in the following way by the act: The principal purpose of the action must be recorded, that is, energy production on the one hand or volume reduction with elimination of pollutant potential on the other hand. The basis for determining the principal purpose is the individual waste, without allowance for its blending with other substances. Furthermore, the nature and extent of its impurities as well as the further wastes and emissions generated by its processing determine whether the principal purpose is oriented toward reclamation or treatment. This means that the larger and more hazardous the quantity of further wastes and emissions, the more a thermal treatment of wastes is indicated. This thermal treatment has no priority over the disposal of household waste, so that incineration of the latter is not mandated.

2.4. Legal Formulation of Recycling

Articles 5 to 9 of the Recycling Act govern the duties of the producer and owner of wastes. The fundamental duties of the consignees are addressed in § 5. These are concretized in the following standards.

Article 5 Sect. 1 concerns waste avoidance; § 5 Sects. 2 – 5 govern waste reclamation. According to § 5 Sect. 1, the duties related thereto are oriented to § 9 together with regulations issued pursuant to §§ 23 and 24. Waste avoidance is not absolutely made a legal duty in the Recycling Act, but must be implemented only in the framework of regulations yet to be issued.

The concern of § 5 Sect. 2 is reclamation in relation to disposal. Reclamation takes priority over disposal except as provided by the act.

Such provisions are found in § 5 Sects. 4 – 6. The act contains three cases in which the priority of reclamation is not applicable: nonfeasibility of fulfilling the criteria cited in Sect. 4; waste disposal as a more environmentally safe solution; research and development privilege.

According to § 5 Sect. 4, the duty of reclamation of wastes must be observed provided this is technically feasible and economically acceptable, in particular, that a market exists or can be created for a substance recovered or energy extracted. According to § 5 Sect. 5, the priority of reclamation of wastes is set aside when disposal of the wastes is the environmentally safer solution. The act establishes four criteria to be used in assessing which solution is environmentally safer: (1) the emissions to be expected, (2) the goal of conserving natural resources, (3) the energy to be consumed or produced, and (4) the enrichment of pollutants in products, wastes for reclamation, or articles obtained therefrom.

The relationship of material and energetic reclamation is governed by § 6. The environmentally safer mode of reclamation takes priority. The criteria of § 5 Sect. 4 apply. The German Federal Government can, by regulation pursuant to § 6 Sect. 1 Clause 2, state the priority of material or energetic reclamation for individual types of substances. If no such regulation exists, § 6 Sect. 2 permits energetic reclamation only if (1) the heating value of the individual waste, without blending with other substances, is at least 11 000 kJ/kg, (2) a combustion system efficiency of at least 75 % is achieved, (3) the heat evolved is itself utilized or conveyed to third parties, and (4) the further wastes generated in the course of reclamation can be placed in landfills, if possible without further treatment. Requirements on the quality of reclamation are also set forth in § 5. According to Sect. 2 Clause 2, a high-quality reclamation consistent with the nature and quality of the wastes is to be sought. The Federal Government can, under § 7, issue regulations governing the details of recycling, in particular of waste reclamation.

2.5. Disposal Standard – TA Abfall

According to § 10 Sect. 4 of the Recycling Act, waste disposal must not impair the welfare of the public. In particular, the health of persons must not be harmed; animals and plants endangered; waters and soil harmfully affected; noxious environmental impacts brought about by air or noise pollution; the interests of regional policy or land-use planning, nature conservation, and landscape conservation, as well as urban development interfered with; or the public safety and order threatened or disturbed. Articles 10 and 11, however, leave the question open as to what technical requirements are to be imposed on the nature and mode of waste disposal in individual cases for reasons of environmental protection. In § 12 Sect. 2, it is therefore provided that such details can be established in general administrative regulations by the Federal Government.

2.5.1. TA Abfall

On the basis of the standard corresponding to this provision (§ 4 Sect. 5 of the Waste Act of 1986), the first part of the Second Technical Directive on Waste (2. Technische Anleitung Abfall) was issued in two subparts [16]. It establishes a new technical framework of conditions for waste treatment, consistent throughout the Federal Republic, and, like TA Luft and TA Lärm (Technical Directives on Air and on Noise, respectively), is binding only on the executive, not the citizen. It guides the discretion of the state authorities having jurisdiction over the execution of the Waste Act and its implementing regulations. The authorities base their technical tests on the standards of TA Abfall. In particular for the approval and operation of hazardous waste disposal facilities and for the allocation of wastes to disposal paths in accordance with §§ 8 ff of the Waste and Residue Monitoring Regulation, TA Abfall contains instructions to the administration. The technical requirements of the part issued in April 1990 concern the storage, physicochemical and biological treatment and combustion of wastes subject to special monitoring. The supplement of December 1990 has to do with their aboveground and underground disposal. Wastes to which TA Abfall is applicable are conclusively enumerated in Annex C. The listing of 346 types of waste subject to special monitoring is identical with the Annex to the Waste Determination Regulation. It contains, however, additional columns for the allocation

of wastes to certain disposal paths, specifically to:

1) Chemical-physical-biological (CPB) treatment facilities
2) Household-waste incineration plants or other waste incineration plants to which the TA Abfall does not apply
3) Incineration plants for wastes subject to special monitoring (hazardous-waste incinerators)
4) Household-waste landfills or other landfills to which the TA Abfall does not apply
5) Aboveground landfills for wastes subject to special monitoring (hazardous-waste landfills)
6) Underground disposal sites in salt formations for wastes subject to special monitoring

In this way, the TA Abfall allocates wastes for the first time to certain disposal paths in order to prevent such wastes from impacting the environment or being deposited in household-waste landfills for lack of alternatives.

Number 4.4 of the TA Abfall governs the criteria under which wastes are to be allocated to disposal methods and facilities. They will have the consequence that wastes must be increasingly refused landfill disposal and must be forwarded to preliminary treatment. This leads to the situation where, for the treatment and preparation of wastes subject to special monitoring for landfill disposal, supplemental treatment capacities must be created on a large scale. Implementation of this allocation begins, according to Number 4.2, with a general ban on mixing, so that the waste composition and the pollutant level are not modified in order to evade the intended allocation.

2.5.2. TA Siedlungsabfall

The Federal Government has issued the Technical Directive on Municipal Waste (TA Siedlungsabfall) [17], which came into force on June 1, 1993. In contrast to the TA Abfall, Part 1, which applies to all wastes cited in the Waste Determination Regulation, a delimitation of the waste types to be included had to be found for the TA Siedlungsabfall. It covers the following waste categories: household waste, bulky waste, residential-type commercial wastes, garden and park wastes, market wastes, street sweepings, construction wastes, sewage sludge, fecal matter, fecal sludge, residues from wastewater-treatment plants, water-treatment sludges, and product-specific wastes, that is, wastes arising in industry, trade, or other institutions that are not municipal wastes but whose nature, pollutant content, and reaction behavior allows them to be disposed of like municipal wastes.

TA Siedlungsabfall is intended to regulate the requirements on the disposal of municipal wastes, consistent throughout the Federal Republic, such that the welfare of the public is not impaired [18]. For this reason, the technical directive contains framework conditions for the disposal of municipal wastes with allowance for disposal safety, resource conservation, relief of load on disposal facilities, and diminution of pollutant impacts on the environment. The goal is to support bodies with the duty of disposal in taking on the task of environmentally safe disposal of municipal wastes to be performed by them. The Federal Government regards a future-oriented, integrated waste management concept to be essential for handling the problem of waste. This means that an interconnected, intercoordinated system of methods and techniques for avoidance, pollutant level reduction, material reclamation, treatment of wastes not lending themselves to material reclamation, and environmentally safe disposal of pretreated wastes should be applied to the various types of municipal wastes.

The new administrative regulation established strict environmental requirements on the disposal of municipal wastes. The goal is to reduce the quantities of municipal wastes reaching the landfill. The Federal Government, in issuing TA Siedlungsabfall, pursued the goal of having the bodies with the duty of disposal largely exhaust the possibility of material reclamation and pollutant level reduction. To this end, the criteria for the allocation of waste streams to disposal processes and requirements on material reclamation (including composting) and pollutant level reduction are standardized. An essential criterion for reducing the amount of waste to be landfilled is reclamation. In order to secure this, TA Siedlungsabfall begins with the principle of segregation. Thus pollutants and valuable substances are always to be collected separately in order to permit utilization of the valu-

able substances or to facilitate the handling of reclaimable residual wastes.

An important part of the technical directive concerns future requirements on environmentally safe landfill disposal of residual wastes that cannot be avoided or reclaimed. Even if all possibilities of meaningful reclamation are exhausted, there will always be municipal wastes that must be permanently disposed of in landfills. To this end, the technical directive sets up requirements, first, on the quality of wastes for landfilling and, second, on the sealing systems necessary in future landfills. For this, the technical directive provides for the pretreatment of wastes to be disposed of in landfills, in particular by incineration. Furthermore, it creates requirements on environmentally safe disposal of residual wastes that cannot be avoided and reclaimed by means of a variety of landfill types.

2.6. Hierarchy of Responsibility

According to § 11 Sect. 1 of the Recycling Act, the producer or owner of wastes has the duty to dispose of them. Anyone who produces or owns wastes, including private persons, himself has the duty of disposal. This statement is negated for certain cases in §§ 13 to 18, but nothing in the fundamental duty is changed. Exceptions to the duty, existing for everyone, to dispose of wastes oneself appear in § 13. The first exception is for wastes from private households. If private households are not in a position to dispose of their wastes or if they have no intention of private disposal, then under § 13 Sect. 1 Clause 1 they must yield these wastes to corporate bodies obligated under state law to dispose of them. This also holds for producers and owners of wastes for disposal from other areas of origin if they do not dispose of these wastes in their own facilities or if overweighing public interests require such yielding. The exception stated in Clause 2 holds for any conceivable waste; the duty to yield must be concretized by a municipal ordinance. According to § 13 Sect. 2, the duty to yield wastes under Sect. 1 does not exist with respect to public-law disposal utilities when the duties of reclamation and disposal have been transferred to third parties or private disposal utilities. The duty to yield wastes does, however, exist under § 13 Sect. 2 Clause 3 for wastes that (1) are subject to a take-back or give-back duty based on a regulation pursuant to § 24, unless the public-law disposal utilities cooperate in the take-back on the basis of a provision pursuant to § 24 Sect. 2 Number 4; that (2) are delivered via public collection to an orderly and innocuous reclamation; that (3) are delivered via commercial collection to an orderly and innocuous reclamation, provided this is demonstrated to the public-law disposal utilities and not overweighing public interests stand against this.

According to § 13 Sect. 4, the German Federal States can (as some already do) decree duties to deliver and yield wastes subject to monitoring.

Corresponding to the duty to yield wastes is the duty of public-law disposal utilities to dispose as set forth in § 15. The duty to reclaim according to the standards of § 5 Sect. 4 takes priority over the duty to dispose.

Under § 16, the public-law disposal utility can transfer its legal duties in whole or in part to a third party, with the consequence that the public-law disposal utility involved is released from all duties. This is, however, possible only under the conditions set forth in detail in the statute.

2.7. Approval of Waste Disposal Facilities

All waste disposal facilities, whether internal to a company or of a public-law nature, are subject to inspection by the state agency having jurisdiction; this provision refers to landfills, refuse incineration and composting and other plants, as well as equipment in which wastes are treated, stored, or transferred.

For both the erection and the operation of a fixed waste disposal facility, § 31 Sect. 1 provides for approval of the plan by the agency having jurisdiction. The sense and purpose of this rule is to afford the agency the possibility of verifying whether the conditions of an environmentally safe waste disposal facility are satisfied. The approval procedure, which is set forth in the Administrative Procedures Act (Verwaltungsverfahrensgesetz, VwVfG), is divided into three phases: creation of the plan by the proposer of the project; the hearing and discussion procedure opened on submission of the plant at the hearing (§ 79 VwVfG), and the procedure involving the establishment of the plan (§ 74

VwVfG). In the best case it offers an assurance that the many interests concerned will be comprehensively examined, weighed, and balanced.

In place of the plan approval procedure, on application or on agency initiative, pursuant to § 31 Sect. 3 of the Recycling Act, an approval procedure can be carried out if the project concerns the erection or operation of an insignificant waste disposal facility, if objections are not anticipated, or if substantial changes in a facility or corporation are at issue. As a result of this possibility of plan approval, the procedure is supposed to be simplified and speeded up. It is, however, of little importance with respect to the disposal of wastes subject to special monitoring, since as a rule a hazardous waste incineration plant or landfill will not be an insignificant facility. Because of the lack of public acceptance, objections can be expected as a matter of course. Thus § 31 Sect. 3 will come into play, at most, with regard to the approval of substantial modifications in an existing facility or its operation.

Even before the establishment of the plan or before granting of final approval, performance can be started in advance under the conditions cited in § 33.

The approving agency is required to approve only those facilities that satisfy the requirements of TA Abfall; that is, they must fulfill the criteria for the allocation of wastes to disposal methods and to disposal facilities, as well as the requirements imposed on the facilities. Annex A of TA Abfall contains details on approval applications and application documents.

Unapproved facilities must either be abandoned and removed or legalized after the fact. In either case, operation of such facilities is an offense under § 327 Sect. 2 No. 2 of the German Criminal Code (Strafgesetzbuch).

As a rule, the approval of a waste disposal facility takes place by plan approval procedures. No other official decisions are required aside from plan approval; in particular, no publiclaw approvals, grants, permissions, consents, endorsements or other plan approvals are required. A plan approval decision governs all public-law relationships between the proposer of the project and the parties concerned. The plan approval decision must include in its finding all considerations of substantial weight. It must leave no open questions [19].

In the framework of a plan approval procedure, the environmental safety of the facility must always be tested pursuant to § 31 Sect. 2 Clause 2. This environmental safety test is a non-self-contained procedural component of the plan approval (§ 2 Sect. 1 of the Environmental Safety Test Act, UVPG). This means that the environmental safety test takes place within the plan approval procedure pursuant to waste law, i.e., there is no separate environmental safety test procedure and no special environmental safety test report. Environmental impacts in the sense of the Environmental Safety Test Act are changes in human health and in the physical, chemical, or biological quality of individual constituents of the environment or the environment as a whole. The testing of these environmental impacts is prescribed as to form, and must follow the procedural steps required by §§ 5 to 12 of the UVPG.

A complaint against a plan approval decision – an administrative action in the sense of § 35 Sect. 1 of the Administrative Procedures Act – can be lodged without prior objection, § 79 VwVfG, in conjunction with § 69 of the Administrative Courts Regulation (VwGO); that is, no prior procedure is required. In order to further speed up the action of law, the "Oberverwaltungsgericht" (a high administrative court) decides in the first instance, pursuant to § 48 Sect. 1 No. 5 VwGO, if there is opposition to the erection, operation, or a substantial modification of a fixed waste disposal facility in which extra wastes are stored or transferred or in which more than 100 000 t/a of wastes are incinerated or thermally degraded.

2.8. Monitoring of Waste Disposal

The numerous provisions of waste legislation can exert their environmentally protective actions only if there is an assurance that their observance is monitored and secured by effective control mechanisms.

2.8.1. Monitoring

According to § 40 Sect. 1 of the Waste Act, all phases of waste disposal are subject to monitoring by the agency having jurisdiction. This means that not only the owners of wastes, the

proprietors of waste disposal facilities, and natural persons and corporate bodies of civil law concerned with the collection and transportation of wastes are subject to monitoring, but also the public-law disposal utilities (pursuant to § 13 of the Recycling Act) and third parties contracted by them (§ 16). Articles 42 and 43 govern the various accountabilities of the owners of wastes. According to § 43 of the Waste Act, producers, collectors, or transporters and disposers of wastes subject to special monitoring are directly obligated by force of statute

1) To keep a book of accounts
2) To present certificates
3) To demonstrate to the agency having jurisdiction that the legal conditions for special accountability are satisfied on their part

How the accounts of the whereabouts of wastes are to be kept pursuant to § 43 of the Waste Act is governed in detail in the Waste and Residue Monitoring Regulation [20], which superseded the Waste Accountability Regulation (Abfallnachweis-Verordnung). According to §§ 8 and 12, an owner of wastes or residues must declare the permissibility of disposal before performing the disposal. In this way, the regulation creates an obligation of prior demonstration, as is already practiced by many landfill operators.

The disposal permit according to Annex 3 is to be used when the disposal of wastes requiring a permit is contemplated, i.e., for extra wastes and other wastes when accountability has been instituted by the agency having jurisdiction over wastes. It is valid for a maximum of five years.

The collection and disposal permit, which can simplify the procedure under the conditions of § 10, comes into play in the collection of small quantities of wastes containing pollutants.

The simplified disposal permit of § 12 is to be used for refuse not subject to accountability.

For household waste and other wastes that can be disposed of with wastes produced in households, no disposal permit is required, even for household waste containing pollutants.

According to § 42 Sect. 1 of the Recycling Act, the agency having jurisdiction can by administrative action obligate the owners of such wastes that cannot be disposed of together with wastes produced in households, or can be so disposed of only under special precautions, to

1) Present evidence of the nature, quantity, and disposal of wastes (general accountability)
2) Keep an accounts book
3) Retain and safeguard certificates and accounts and submit certificates for examination (special accountability)

Such an arrangement is within the discretion of the regulatory agency. It can therefore institute the permit only for certain wastes or solely with respect to a means of accountability.

2.8.2. Transportation Approval

In order to ensure orderly waste disposal, it is further necessary that the authorities be able to examine the specialist, technical, and personal conditions of those persons and companies that collect and transport wastes. Here § 49 of the Recycling Act offers a legal foundation that provides a transportation approval for companies that commercially collect wastes for disposal. Unsuitable persons or companies can thus be prevented from handling wastes in dangerous or incorrect ways.

Commercial collection and transportation takes place when there is a self-contained activity aimed at earning a profit and created for a certain span of time. It takes place in the framework of commercial enterprises when it is performed with no profit motive for internal purposes of a commercial company or a public enterprise producing goods for consumption. The latter, practically, is the case only when the company brings its own wastes to the waste disposal facility, so that the owner of the wastes is simultaneously also the transporter of the wastes.

The form of the approval procedure is governed by §§ 4 – 7. In order to grant a transportation approval, the authorities test the reliability of the applicant and, for this purpose, request the documents listed in § 4 Sects. 1 and 2. For this reason, the approval is not transferrable (§ 6). Reliability presupposes that the transporter employs personnel subject to his instructions. Further, the capability of the waste transporter must be verified (§ 4 Sect. 1). The transporter must demonstrate adequate liability insurance, with a limit of at least € 2.6×10^6. Transportation approval may be denied on the grounds of the general welfare if the application is for roundabout transportation with an excessively high risk.

If all the conditions are met, in particular if there is no anxiety as to impairment of the public welfare, the applicant has a claim to issuance of the approval pursuant to § 49 Sect. 1 Clause 1 of the Recycling Act.

2.9. Plant Officer for Waste

Official monitoring of waste disposal is supplemented by internal monitoring. For facilities in which wastes subject to special monitoring are produced regularly, and also for fixed waste disposal facilities, a plant officer for waste must be designated. Which plants are affected by this provision is set forth in detail in § 54 Sect. 1 of the Recycling Act and § 1 of the Regulation on the Plant Officer for Waste (Verordnung über den Betriebsbeauftragten für Abfall, AbfBetrbV).

Essentially, a plant employee must be designated as the plant officer for waste. This person's knowledge of the operation should help reveal opportunities for waste avoidance and waste reclamation. A person from outside the plant may, however, also be appointed plant officer. In this case, prior agreement of the agency having authority is required under § 4 of the AbfBetrbV.

The tasks and authority of the plant officer for waste are set forth in § 55 of the Recycling Act. This officer must observe, examine, and collaborate in the actions identified in § 55 Sect. 1 No. 1 to 5 of the Waste Act, must submit reports, must issue a position on certain investment decisions, and must advise, report, and cooperate. Within the plant, he must aid in the performance of duties under waste law, for example in waste avoidance and reclamation, and in particular must advise the plant owner in these matters. In his annual report, the plant officer must report in what monitoring actions and inspections have been performed and whether any reports of plant upsets have been submitted.

2.10. Waste Management Concepts and Balances

2.10.1. Waste Management Concepts

According to § 19 of the Recycling Act, producers generating in all 2000 kg/a of wastes subject to special monitoring, or more than 2000 t/a of wastes subject to monitoring per waste code, must create a waste management concept relating to the avoidance, reclamation, and disposal of the wastes produced. The producer's waste management concept functions, on the one hand, as an internal planning tool and, on the other hand, as a basis for official waste management planning. To this end, it must be presented for evaluation if requested by the agency, having jurisdiction. The content of a waste management concept is governed in detail by § 19 Sect. 1 Clause 3. According to Sect. 2, the waste management plan data pursuant to § 29 must be taken into consideration when this plan is prepared.

2.10.2. Waste Management Balances

According to § 20 Sect. 1, those having duties in the sense of § 19 Sect. 1 must prepare annually (starting on April 1, 1998) a balance for the prior year, showing the nature, quantity, and whereabouts of wastes subject to special monitoring and wastes subject to monitoring that have been reclaimed or disposed of. According to § 20 Sect. 2, those with the duty of preparing a waste balance have a right to information relative to owners of wastes from commercial or other economic enterprises or public institutions insofar as these have yielded wastes to those having the duty.

2.11. Remediation of Hazardous Waste Sites

In order to avert hazards to human health and groundwater quality, hazardous waste sites must be remediated. The legal conditions for such remediation have not yet been formulated in a self-contained way applicable throughout the Federal Republic. Rules for the identification, testing, hazard evaluation, and remediation of hazardous waste sites as well as financing are found in the bodies of law relating to wastes, water, mining, police powers, and order.

2.11.1. Duty to Remediate

Since the Waste Disposal Act (Abfallbeseitigungsgesetz) of 1972 took effect, the creation

of new hazardous waste sites with unforeseeable financial consequences for the public has been barred. According to § 36 of the Recycling Act, the waste authorities, under § 36 Sect. 2 of the Recycling Act in the case of the closure of a waste disposal facility, and under § 36 Sect. 3 of the Waste Act in the case of the closure of commercial facilities in which wastes subject to special monitoring are produced, can require recultivation as well as protective measures if this is necessary to guard against impairment of the public welfare. The operator of the facility and the proprietor of the land have the duty of remediation.

For conditions dating from the time from March 1, 1960 [when the Water Resources Management Act (Wasserhaushaltsgesetz) took effect] until June 11, 1992, the Water Resources Management Act can be applicable if the storage or landfill disposal of substances fulfills a disallowed water use (§ 3 Sect. 1 No. 4 or 5, § 3 Sect. 2 No. 2, § 22 Sect. 2, § 26 Sects. 1 and 2, § 34 Sect. 2, all of the Water Resources Management Act). Substances may be stored near a body of water only in such fashion that there need be no concern over harmful contamination of the water or any other detrimental change in its properties or in the discharge. On the basis of the framework provisions of the Water Resources Management Act and the state water legislation supplementing it, the authorities can order remediation [21]. Liability for costs is governed by state legislation.

For facilities closed or contaminations caused before March 1, 1960, only the state water legislation or its predecessor legislation applies insofar as it contains provisions extending beyond the general police power. Because of the constitutional ban on a genuine retroactive character of statutes, there is no recultivation and remediation duty under German Federal Law extending back beyond March 1, 1960.

Subsidiarily, the provisions of general police power are cited for averting hazards from hazardous waste sites. The police power is, however, oriented toward the maintenance of present safety and order; it permits necessary actions to limit damage but, as a rule, not remediation actions. A hazard originating from a hazardous waste site can be responded to under police powers by recourse to the polluter or also to the proprietor or possessor of actual power.

2.11.2. Other Duties

For all waste disposal facilities that were operated or erected before June 11, 1972 and are still in service, the agency with jurisdiction can, pursuant to § 35 of the Waste Act, issue subsequent orders, set deadlines, conditions, and specifications, and even order the cessation of operation. For facilities approved after June 11, 1972, changes in operation can be ordered only pursuant to § 32 Sect. 4 Clause 2 of the Recycling Act.

Under § 40 Sect. 1 Clause 2 of the Recycling Act, monitoring under waste law can also be extended to closed waste disposal facilities and to plots of land on which wastes were produced, treated, stored, or transferred before June 11, 1972, provided this is necessary to protect the public welfare.

If dwellings have been built on hazardous waste sites, the agency with jurisdiction can order their demolition if the hazardous waste sites present health risks to the residents.

2.11.3. Financing of Remediation of Hazardous Waste Sites

The remediation of hazardous waste sites entails high costs. A reliable total cost estimate does not exist, but a figure spoken of is up to € 35.8×10^9. Various financing models are under discussion and have already been introduced in some German Federal States.

The License Model of North Rhine-Westphalia. The North Rhine-Westphalia State Waste Act, in its fourth part, governs the issuance of licenses for the treatment or transfer of wastes that are excluded under § 15 Sect. 3 of the Recycling Act (§ 10 Sect. 1 of the North Rhine-Westphalia State Waste Act). Licenses are granted by the State Water and Waste Office. According to § 11 Sect. 1 of the North Rhine-Westphalia State Waste Act, a fee is charged for working the license; by § 15 Sect. 1 of the North Rhine-Westphalia State Waste Act these fees are dedicated to hazardous waste site remediation and further tasks of the "Entsorgungsverband" or disposal association. This model is intended

to solve problems of hazardous waste disposal and hazardous waste site remediation at the same time. The coordination of both areas by means of the dedicated fees and the mandatory membership of private disposal companies in the North Rhine-Westphalia Disposal Association (§ 5 of the Disposal Association Act) is justified with numerous parallels that can lead to a transfer of technology and know-how.

The Cooperation Model. In Rhineland-Palatinate and Hesse, the "cooperation" model is in effect. The state and local governments and businesses cooperate in administering hazardous waste site remediation by establishing a company to dispose of hazardous wastes with remediation functions.

2.12. Criminal Offenses and Infringements of Regulations

Unlawful disposal of waste is punishable, under numerous provisions of German Federal and State Law, by monetary penalties and fines and, in some cases, by criminal fines or imprisonment.

German Federal Law provides criminal penalties for offenses as follows:

– § 326 of the Criminal Code (waste disposal endangering the environment)
– § 327 Sect. 2 No. 2 and Sect. 3 No. 2 of the Criminal Code (unapproved operation of a waste disposal facility)
– § 330 (severe endangerment of the environment)
– § 13 in conjunction with § 7 Sect. 3 Clause 2 of the Radiation Protection Precautions Act (violation of dose and contamination values)

Fines are assessed for the following offenses:

– § 61 of the Recycling Act (unlawful waste disposal)
– § 7 of the AbfKlärV (unlawful handling of sewage sludge)
– § 9 of the Waste Accountability Regulation (incorrect bookkeeping)
– § 5 of the PfandV (failure to levy deposit on plastic bottles)
– § 12 of the Packaging Regulation (packagings harmful to the environment)
– § 10 of the Waste Oil Regulation (unlawful handling of waste oil)
– § 19 Sect. 1 No. 2 of the Dead Animal Disposal Act (unlawful disposal of dead animals or animal body parts other than in dead animal disposal facilities)
– § 46 Sect. 1 No. 4 of the Atomic Energy Act (unlawful disposal of radioactive substances)
– § 41 Sect. 1 No. 9 in conjunction with § 26 Sect. 1 of the Water Resources Management Act (release of solids into a body of water for the purpose of disposing of them)

What is understood by "waste" in criminal law is based on the waste concept of the Waste Act, but is oriented to different standards than in waste law. What is at issue is not the intent of the owner of movable property to reclaim it, but its objective use value. If, for example, an item of property is at present objectively without use value because of its contamination, but its condition is such as to threaten the environment, then it is waste in the sense of § 326 of the Criminal Code. This holds even if the owner wishes to reuse or reclaim the substance or its constituents after disposal [22]. The waste concept of § 326 of the Criminal Code also differs from that of waste law in that the criminal law does not adopt the application restrictions provided in § 1 Sect. 2 of the Waste Act for substances whose disposal is governed by special provisions. On the other hand, the criminal concept of waste is not enlarged by the fictions of §§ 5 Sect. 2, 5a, and 15 of the Waste Act.

2.13. Civil Liability

2.13.1. Product Responsibility

According to § 22 Sect. 1 of the Recycling Act, product responsibility for fulfilling the goals of recycling is borne by the one that develops, manufactures, finishes, and processes or sells products. For the fulfillment of product responsibility, products must be designed so that the production of wastes is decreased as much as possible in their manufacture and use, and the environmentally safe reclamation and disposal of wastes produced after their use is ensured. Product responsibility is defined more closely

in § 22 Sect. 2 of the Recycling Act. Who bears product responsibility is not based on the statute; instead, the German Federal Government determines, by regulation pursuant to §§ 23 and 24 of the Recycling Act, what parties having duties are to fulfill the product responsibility according to § 22 Sects. 1 and 2 of the Recycling Act. The Federal Government further established by regulation for what products and in what fashion and manner product responsibility is to be assumed.

2.13.2. Damage Claims

The Recycling and Waste Act contains no provisions as to the civil liability for waste-related harm that may be done on the path from waste production to final waste disposal. Claims for compensation of damages are furthermore governed by the general rules.

Noncontractual liability for damages caused by wastes is essentially determined according to the provisions of §§ 823 to 853 of the German Civil Code. Anyone harming the life, the body, the health, the freedom, the property, or any other right of another by malice or negligence is obligated to compensate the other for the damage. The damage must be caused illegally and culpably. The jurisprudence here has developed on the basis of a producer liability, which extends so far that even wastes in the field of industrial wastes are included in the liability and a responsibility for the state/safety of something ("Verkehrssicherheitspflicht") has been developed and facilitations of the burden of proof ("Beweiserleichterungen") have been created. According to § 823 Sect. 2 of the Civil Code, furthermore, violation of a statute that seeks to protect another creates an obligation of compensation for damages. Article 22 of the Recycling Act in conjunction with regulations still to be issued pursuant to §§ 23 and 24 of the Recycling Act is such a protective statute. Claims for compensation of damages can also arise from violations of contractual duties. The basis of the claim is then the institute of positive infringement of claim [23] as developed by jurisprudence. The Product Liability Act [24] was adopted to protect the consumer against harm to his health and his property by a defective product. This statute justifies a liability independent of culpability, which can also come into the picture when a defective product causes a harm upon its disposal. The Environmental Liability Act [25] introduces a liability of endangerment, independent of culpability, for the operators of one of the facilities enumerated in the Annex to the act. In this way, the safety of these facilities is supposed to be enhanced. If the environmental impact of a facility kills a person, harms another, or damages property, the proprietor of the facility must make recompense for damages.

3. Organization of Hazardous-Waste Disposal

At present the catalog of hazardous wastes includes 331 well-defined substances; presumably the number will increase as more and more substances are recognized as meriting special attention.

The quantity of hazardous wastes is decreasing because waste-prevention and waste-reduction efforts in industry and business, such as closed process loops for acids and scrubbing liquors, are succeeding; waste reclamation is being increasingly practiced in waste-generating plants; and because recycling processes have been improved.

With regard to quality, too, the generation of hazardous wastes will continue to change. Because pollutant concentrations in hazardous wastes are increasing, the quality of treatment has to meet more stringent requirements. Wastes for deposition must be subjected to increasingly differentiated pretreatment.

As a consequence, more and more effort and expense are going into the organizational, technical, and cost aspects of hazardous-waste treatment. Logistics as well as treatment and reclamation management must be flexible so as to adapt to waste quality and quantity, regional variations in occurrence, and technical developments. Requirements on the organization of hazardous-waste management must be viewed against this background.

The German hazardous-waste disposal market is split in two. The large chemical companies have their own high-quality disposal facilities, which, with a few exceptions, are not available for waste disposal by other companies. These industrial and commercial businesses transfer

Table 7. Hazardous-waste disposal companies in Germany [26]

Land	Address		Share capital, €	Ownership
Baden-Württemberg	SBW	Sonderabfallentsorgung Baden-Württemberg GmbH, Welfenstr. 15, 70736 Fellbach, Tel.: 0711/51839-0	25 053 302	99.00 % Baden-Württemberg State 1.00 % Council of Cities and Rural Districts
Bavaria	GSB	Gesellschaft zur Entsorgung von Sondermüll in Bayern mbH, Winzerstr. 47d, 80797 München, Tel.: 089/30629-0	32 104 017	78.84 % Bavaria State 10.03 % 74 private sector companies 8.60 % Top-level communal associations 2.53 % Bavarian state institution for construction financing
	ZVSMM	Zweckverband Sondermüllentsorgung Mittelfranken, Siemensstr. 3 – 5, 91126 Rednitzhembach-Igelsdorf, Tel.: 09122/797-0	15 338 756	100 % 16 communes in Middle Franconia administrative region
Berlin/Brandenburg	SBB	Sonderabfallgesellschaft Brandenburg/Berlin mbH, Berliner Straße 50, 14467 Potsdam, Tel.: 0331/2708873 (relocation scheduled from May 1995)	1 533 875	25.00 % City of Berlin 25.00 % Brandenburg State 25.00 % GbR der Erzeuger (20 waste-generating companies and associations) 25.00 % GbR der Entsorger (26 disposal companies)
Bremen	BSBG	Bremer Sonderabfallberatungsgesellschaft mbH, Knochenhauerstr. 36/37, 28195 Bremen		51.00 % City of Bremen 49.00 % Private sector ownership
Hesse	HIM	Hessische Industriemüll GmbH, Kreuzberger Ring 58, 65205 Wiesbaden, Tel.: 0611/7149-0	29 549 000	55.00 % Private sector ownership 28.00 % Hesse State 27.00 % Recon GmbH/Bayernwerk AG
Lower Saxony	NGS	Niedersächsische Gesellschaft zur Endablagerung von Sonderabfall mbH, Alexanderstr. 4/5, 30159 Hannover, Tel.: 0511/3608-0	10 225 837	51.00 % Lower Saxony State 13.00 % Preussag AG 13.00 % Salzgitter AG 12.85 % Disposal companies 8.50 % Waste-generating companies 1.65 % Hannoversche Beteiligungsgesellschaft
Rhineland-Palatinate	SAM	Sonderabfall-Management-Gesellschaft Rheinland-Pfalz mbH, WilhelmTheodor-Römheld-Str. 34, 55130 Mainz, Tel.: 06131/98298-0		51.00 % Rhineland-Palatinate State 49.00 % Disposal businesses
Saarland	SES	Sonderabfallentsorgung Saar GmbH, Ursulinenstr. 35, 66111 Saarbrücken, Tel.: 0681/38704-0	3 500 000	49.00 % Saarberg-Oekotechnik 25.90 % Vereinigte Saar-Elektrizitäts AG 25.10 % Saarland State
Schleswig-Holstein	GOES	Gesellschaft für die Organisation der Entsorgung von Sonderabfällen, z. Zt. c/o Minister für Natur und Umwelt, Grenzstr. 1 – 5, 24149 Kiel, Tel.: 0431/219356		25.90 % Schleswig-Holstein State 25.10 % Districts and cities not contained in districts 24.50 % Private waste-generating companies 24.50 % Private waste-disposal companies
Thuringia	TSA	Thüringer Sonderabfallgesellschaft mbH, Waidmühlenstr. 1, 99102 Waltersleben, Tel.: 0361/6437-0	15 000 000	75.00 % Thuringia State 12.50 % Hessische Industriemüll (HIM) 12.50 % GSB (Bavaria)

Table 8. Hazardous-waste organization in the Länder of the Federal Republic of Germany [26]

State	Obligation to offer service?	State company [a]	Have own disposal facilities?			Have own waste catalog?	Hazardous waste levy?	Private companies allowed?
			PCB	Incineration	Landfill			
Baden-Württemberg	Yes [b]	SBW			X	Yes [h]	Yes	Yes (still)
Bavaria	Yes	GSB		Yes [g]		No	No	No
		ZVSMM		Yes [g]				
Berlin/Brandenburg	Yes [c]	SSB		No		Yes [h]	No	Yes
Bremen	No	BSBG		No information		No	Yes	Yes
Hamburg	No	No		No information		No	No	Yes
Hesse	Yes	HIM	X	X		Yes [i]	Yes	No
Mecklenburg-West Pomerania	No	No		No information		No	No	Yes
Lower Saxony	Yes	NGS			X	Yes [j]	Yes	No
North Rhine-Westphalia	No [d]	No		No		Yes [k]	No [m]	Yes
Rhineland-Palatinate	Yes	SAM		No		Yes [h]	No	No
Saarland	Yes	SES	X			No	No	No
Saxony	No	No		No		No	No	
Saxony-Anhalt	No [e]	No		No information		No [l]	No	Yes
Schleswig-Holstein	No [f]	GOES		No		No	No	Yes (still)
Thuringia	Yes	TSA			X	No	No	No

[a] See Table 7. [b] For hazardous wastes that can be deposited at SAD Billigheim; obligation to offer disposal services is successively extended to more forms of hazardous waste if safe disposal options exist for these substances. [c] Generator and carrier liable for costs with disposal certificate. [d] But disposal autarky takes priority. [e] Not at present, but plans being made with disposal route SAV at least for hazardous wastes. [f] But obligation to advertise and offer services for medical wastes lies with Krankenhaus-Abfallbeseitigung Schleswig-Holstein GmbH, Edisonstrasse 13, 24145 Kiel. [g] No further information. [h] For wastes such that disposal services must be offered. [i] For wastes subject to levy. [j] For waste collection; definition of hazardous wastes goes beyond Waste Definition Regulation. [k] For wastes subject to license. [l] Plans being made for wastes with disposal route SAV. [m] License fees paid by hazardous-waste generators to AAV Abfallentsorgungs- und Altlastensanierungsverband Nordrhein-Westfalen, Werkstrasse 15, 45527 Hattingen.

their especially hazardous wastes to third parties. Often these markets are regulated, and in each German Land the market is dominated by one disposal company. Information on the situations in the Länder can be found in Tables 8 and 7.

In six of the Länder, hazardous-waste disposal is wholly or mainly in public hands because the company is majority publicly owned. In four Länder, state-owned companies handle part of the hazardous waste, and in six this waste is disposed of by private companies (Table 8).

The state companies have the following primary tasks:

1) To ensure the disposal of hazardous wastes by creating and committing adequate disposal capacities
2) To guarantee the proper disposal of hazardous wastes through
   a) Monitoring (waybill system)
   b) Directing wastes to suitable facilities
   c) Maintaining records of waste quantities(statistical recording)

Along with these functions, some states have developed corporate structures and gone into competition with private companies. The associations of private waste-disposal contractors are critical of the situation in those states where such companies have taken over part or all of hazardous-waste disposal, citing:

1) Excess capacity of facilities
2) Lack of flexibility to adapt to changes in waste quantity
3) Price autonomy due to self-created monopoly structure
4) Rising costs, for example due to increased administrative fees, hazardous-waste levies, license fees, and so forth
5) Assumption of consulting functions
6) Elimination of competition in hazardous-waste disposal
7) Harm to competitiveness of industrial regions

8) Impairment of Germany as an industrial location
9) Covert supplemental financing of Land economies

Private disposal contractors also claim superiority over state-owned companies with regard to efficiency. For example, state-owned companies achieve a recycling rate of only 20 %, while private companies had a rate of over 30 % in 1994.

In particular, it seems questionable when state-owned companies mingle governmental functions, such as taxation, supervision, and monitoring, with commercial interests. They control markets and override market principles.

A change also appears called for in the preference for disposal in the same Land. Because enhanced environmental awareness has reduced waste volumes, environmental protection has become integrated into corporations, production processes generating less waste have been adopted, and feedstocks with lower pollutant concentrations have come into use, virtually all reclamation, treatment, and disposal facilities have developed excess capacity. If the existing facilities can accept waste from anywhere in Germany, they will be better utilized, fewer new facilities will have to be built (especially in the former Democratic Republic), and costs for hazardous-waste disposal will decline.

European facilities must also be included in the disposal scheme to the extent that these facilities are as good as or better than domestic facilities and comply with the provisions of the 17th German Federal Pollution Control Act.

It is also important that waste-disposal companies have their waste-treatment facilities certified, because this quality attribute satisfies a security need on the part of their customers.

4. Life-Cycle Assessment

4.1. Introduction

Human activity in the form of production, consumption, and disposal imposes a variety of burdens on and, in some cases, irreversible damage to the environment. The purpose of "sustainable development" is to minimize these environmental burdens. The diverse interrelations between material and energy streams and processing operations makes it uncertain whether individual measures taken together lead to a lessening of such burdens. Accordingly, a method is needed by which environmental burdens due to systems, as well as their mitigation, can be quantified. This method is known as life-cycle assessment (LCA) [27], [28].

Definition and Goal. Life-cycle assessment is a method by which environmental burdens caused by a material or immaterial object of investigation, including potential effects, can be determined. Objects of investigation are, firstly, products and services; companies, facilities, disposal routes, and so forth can also be treated in this way. The LCA consists of the following phases: goal definition and scoping, inventory analysis, impact assessment, and interpretation. In the goal definition phase, reference quantities, system boundaries, and level of detail are set for the assessment as a whole. The inventory analysis phase includes material and energy streams treated as inputs and outputs over the entire life cycle of the products, including the production and preparation of raw materials, production and finishing, distribution, transportation, energy production, consumption and wear and tear, and disposal (including recycling). In the framework of the impact assessment phase, the first two subphases together – classification and characterization – entail a determination of potential environmental impacts (greenhouse effect, ozone degradation, etc.) for each item in the inventory analysis. In the last subphase of impact assessment, interpretation, potential impacts in the various environmental categories are collectively evaluated. The results of a life-cycle assessment are used in selecting the ecologically most favorable variants, environmental optimization of existing systems, and development of new systems (products, services, etc.; see Section 4.2). For this reason, the LCA must be clear, reproducible, and mathematically consistent, and it must be prepared in time periods in accord with knowledge interests [27–31].

Status of the Life-Cycle-Assessment Method. The LCA method is still under development. Inventory analysis has been developed furthest, while interpretation is least advanced.

The technique of classification and characterization occupies an intermediate position. While both inventory analysis and the classification and characterization subphases are based on scientific rules, this is only partly possible in the interpretation subphase, which involves environmental objectives among other factors [29].

In spite of the partly subjective valuation technique and other fuzzy areas, LCA is a very good decision aid in industry and commerce, since the most complex relationships are systematically reduced to a few data sets. LCA is already a useful instrument for objectivizing the environmental burdens that actually result from an object of investigation. Life-cycle assessments must be performed under recognized rules, such as those created by the DIN body NAGUS (Standards Committee on Principles of Environmental Protection) and ISO:

ISO 14 040: Environmental management – life-cycle assessment – principles and guidelines
ISO 14 041: Environmental management – life cycle assessment – life-cycle-inventory analysis
ISO 14 042: Environmental management – life cycle assessment – life-cycle-impact assessment
ISO 14 043: Environmental management – life-cycle assessment – life-cycle-improvement assessment

4.2. Scope and Limits of Life-Cycle Assessments

Life-cycle assessments consider only environmental aspects; they accordingly form just one element needed for decision-making. Technical and safety aspects are taken into account in goal definition; fulfillment of these criteria is essential for the creation of an LCA (Fig. 6). The environmental aspects are the object of the LCA; in contrast, economic and social aspects are additive, built up on the results of the LCA, rather than integral. Product-line analyses have not only environmental but also economic and social aspects as integral components, and this is where they differ from the life-cycle assessment [33].

With the help of the LCA, the environmental relevance of objects of investigation (such as products and services) can be determined in a clear and reproducible way. This accordingly is the basis for ecological optimization; the selection of ecologically "better" alternative products, substances, processes, and designs; and the environmental design and development of products [27], [28], [31].

Table 9 presents applications of LCA as they relate to products. The cases of one and more than one product, and of existing and newly developed products are distinguished. The actions listed serve to minimize the use-related environmental burdens, determined by means of the LCA. For one existing product, life-cycle assessment is employed to optimize environmental relevance over the life cycle by substitution of substances, energies, and processes. For one new product, environmentally compatible development involves the selection of suitable substances, processes, and design. In the case of several existing products, the product having the lowest environmental burden is identified by means of LCA. Any life-cycle assessment is a comparison between objects of investigation, and the same holds for optimization of a product with substance, energy, and/or process alternatives.

Table 9. Scope of application of LCA to products [34]

Product(s)	Existing	New
One	optimization	development optimization
Several (same goal function)	selection optimization	development optimization selection

In addition, the LCA technique can be used to determine the following [27], [28], [31], [35], [36]:

1) The ecologically optimal disposal route for used products and wastes
2) The ecological breakeven point for recycling and other environmental protection practices
3) Weak points in the life cycle (product line) of the object of investigation
4) Priorities of necessary actions

The LCA method is applicable to companies and plants as well as products and services [34], [37]. For each, the method is used to find a local minimum of environmental burdens. These minima, because they do not take account of the full life cycle, do not necessarily lead to a global minimum of environmental burden due to the product.

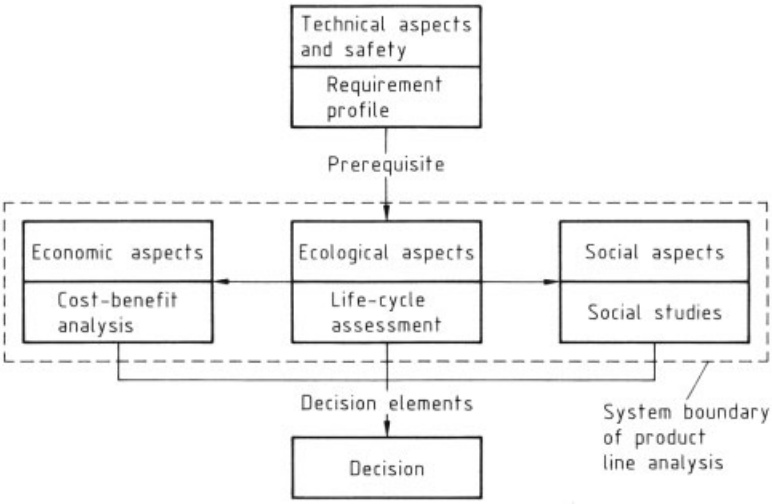

Figure 6. Aspects and elements to be considered in decision-making

The use of LCA has the following limitations [38]:

1) Environmental burdens due to an object of investigation can be determined only relatively, not absolutely, because the true impacts on the environment cannot be determined
2) No universally valid inferences can be made about the environmental compatibility of materials, auxiliaries, and media, since neither the intended use nor the anti-cipated benefit is known. Only the environmental burden up to the manufacture of the substance can be determined as an "ecological mortgage" [34]
3) Universal statements about disposal routes cannot be derived because the reference values for the used product or waste must be known

A life-cycle assessment does not yield unequivocal results (yes/no, black/white, good/bad, . . .) and therefore cannot dictate decisions, merely aid in making them.

An environmentally oriented decision must not be based solely on the result of the LCA, but must also consider the way in which the result was reached; that is, the methodology and the data used are both crucial.

4.3. Method of Product Life-Cycle Assessment

The technique of product LCA can essentially be extended to the other objects of investigation mentioned in the preceding sections. The LCA must be methodologically correct, internally consistent, clear, and reproducible, and it must be generated in the time intervals corresponding to the knowledge interests. Furthermore, it must be done in such a way that the relative environmental relevances of several products, or of one product for which alternatives (substances, energy, processes) are being evaluated, can be determined. The LCA must be in accord with the goal definition.

4.3.1. Goal Definition and Scope

4.3.1.1. Knowledge Interest and System Description

It should be stated at the outset why the LCA is being prepared, for what purpose, and by whom the results and the inferences are to be used [28]. An important question is whether the LCA is internal or public. In an internal life-cycle assessment, it may be appropriate to omit the interpretation step. The internal versus public definition

also governs what kind of critical review (Section 4.4) is to be performed and what personnel will be involved. The level of detail of the LCA must also be identified. The simplest LCA, carried out at the lowest cost, is "screening" in contrast to a "complete" LCA with its high degree of detail.

The level of detail is implied by the system description in the goal definition. The description of the product system and the definition of system boundaries of the LCA must be distinguished. The system to be investigated depends on the concrete knowledge interest. The product itself, the requirements it must meet, the conditions of its use, and other constraints must be explicitly stated. This can be done in the form of a specification. The assessment space is the geographic system boundary of the LCA. It is a delimitable space whose state can be described together with the factors influencing it. The assessment interval is the time for which the assessment is performed. It should be as close as possible to the interval in which the study is carried out; for reasons of symmetry, all data used must originate in the same time interval. The fourth system boundary of the LCA is a methodological constituent of the LCA. The assessment boundaries define the part of the product line that is to be examined and separate it from the surroundings. The assessment boundaries are determined by cutoff criteria, which in turn depend on the level of detail specified in the goal definition and may be stated in terms of a percentage. If the substance or energy in question, relative to the functional unit, falls below a specified percentage, the process chain leading to the creation of this substance or energy is no longer considered as originating in the object of investigation. The procedure for determining the assessment boundaries is iterative and can only be carried out while the inventory analysis is being performed [28], [31].

It must be stated whether the life-cycle assessment is to be carried out for a product line with specific sites taken into account ("specific" LCA) or is to be a "general" LCA based on average values or literature data. In the case of a concrete product line, the sites of production, consumption, and disposal throughout the product life cycle must be identified. The results of both types of LCA and their potential utilities differ depending on these decisions. Mixed forms are also possible, provided data symmetry is guaranteed [39]. With regard to the knowledge interest, the goal definition must also include data-quality targets (accuracy, depth of detail, completeness, consistency, congruency, symmetry, and representativeness) [40].

It must be determined whether the LCA is to be used as a static model for one-time comparison or is to form the basis for optimization and thus has to be used in calculating alternatives and scenarios and in performing sensitivity analyses. For static LCAs, the "black box" method for process assessment is adequate; in this technique, the dependences of the process output on the input are not considered. For nonstatic LCAs, on the other hand, process modeling using these dependences (Section 4.3.2.4.2) is essential [41].

4.3.1.2. Functional Unit

The object of investigation is the computational reference quantity of the product line and represents its primary benefit. All environmental burdens, impact potentials, and additional benefits are referred to the object of investigation, i.e., to the product in the case of a product LCA. Products (finished products) can be compared only if they meet identical needs and satisfy identical requirements (profile of requirements, specification). The functional unit [27], [28] goes beyond the object of investigation per se and, for the sake of comparability between product lines, considers the total benefit furnished by the system (product line) under examination, i.e., both primary and secondary (additional) benefits. The functional unit includes the product lifetime (service life, frequency of use, etc.) and, along with the object of investigation, includes the equality of benefits [42], i.e., the equality of all products created by the systems with respect to quality and quantity (Section 4.3.2.5.4).

4.3.2. Inventory Analysis

Once the goal definition (Section 4.3.1) has been completed and it has been verified that the product as well as the alternatives under consideration fulfill the functional (technical) aspects and

safety requirements according to the specification, the inventory analysis can be undertaken [31].

4.3.2.1. Components of the Product Line

In the context of life-cycle assessment, a process forms the smallest assessable element of an inventory analysis whose inputs and outputs can be identified [41]. The goal of production processes is the process product, while the goal of disposal processes is either the treatment and storage of the waste generated (disposal processes) or the reclamation of used products and wastes (recycling processes). The process goal is the process reference quantity. The product (object of investigation) and the functional unit (Section 4.3.1.2) must be distinguished from the process reference quantity. The process being assessed is broken down into production of raw materials and energy; transportation; production; consumption; and disposal, including recycling. Processes can be linked into process chains, extending up to complete product lines (Fig. 7). The product line is the model of the complex linkage of all processes via the material and energy streams for a finished product within the assessment boundaries, from raw materials extraction through the product to the reincorporation of substances in the environment. This is the model of the product life cycle. Process chains are processes directly linked together through the material stream, often within corporate boundaries, which thus represent parts of the product line. A module consists of a process or process chains that can, by virtue of their unambiguous description and well-defined interfaces, be used for a variety of LCAs [31].

4.3.2.2. Development of the Product Line

The development of the product line has the purpose of identifying all substances, energies, and processes (and process sequences) directly or indirectly involved in the manufacturing, consumption, and disposal of the product. The procedure is as follows [31]:

1) Identify the substances composing the product or that are to compose a new product.
2) Set up the basic structure (main process flow) for these substances, starting from the finished product and going up the material stream back to extraction from the Earth's crust or the biosphere.
3) Prepare the basic structure, starting from the finished product and going in the material stream direction through product use, recycling or waste processing, all the way to landfill disposal of wastes.
4) Determine the input and output substances and forms of energy for each process included in the basic structure.
5) Work out the process chains for each input and output substance and each form of energy identified in the preceding step. The result is the first-order side structure (side process flow).
6) Using the first- and higher-order side structures, proceed as in the fourth step to go from basic structure to the next order of side structure.

Because the assessment boundaries are not yet determined, the development of the product line is broken off after the first-order (sometimes the second-order) side structure has been generated, to limit the amount of work involved. The final development of the product line within the assessment boundaries can be carried out only when data acquisition is complete and the methodological cutoff criteria have been applied.

One of the prerequisites for comparing two or more products having the same purpose is the use of the same cutoff criteria for all products included (on grounds of symmetry) [39].

4.3.2.3. Data Acquisition

Data may be measured, calculated, or estimated; averages or ranges can also be used. A further distinction is whether the data are reproduced in anonymized form or can be associated with a company. For reasons of symmetry, data quality and data type should be consistent within a life-cycle assessment (see Section 4.3.1.1). The origin of data must always be documented. As a prerequisite for the LCA, input and output data must be acquired for each process within the assessment boundaries. Data acquisition begins with the process description.

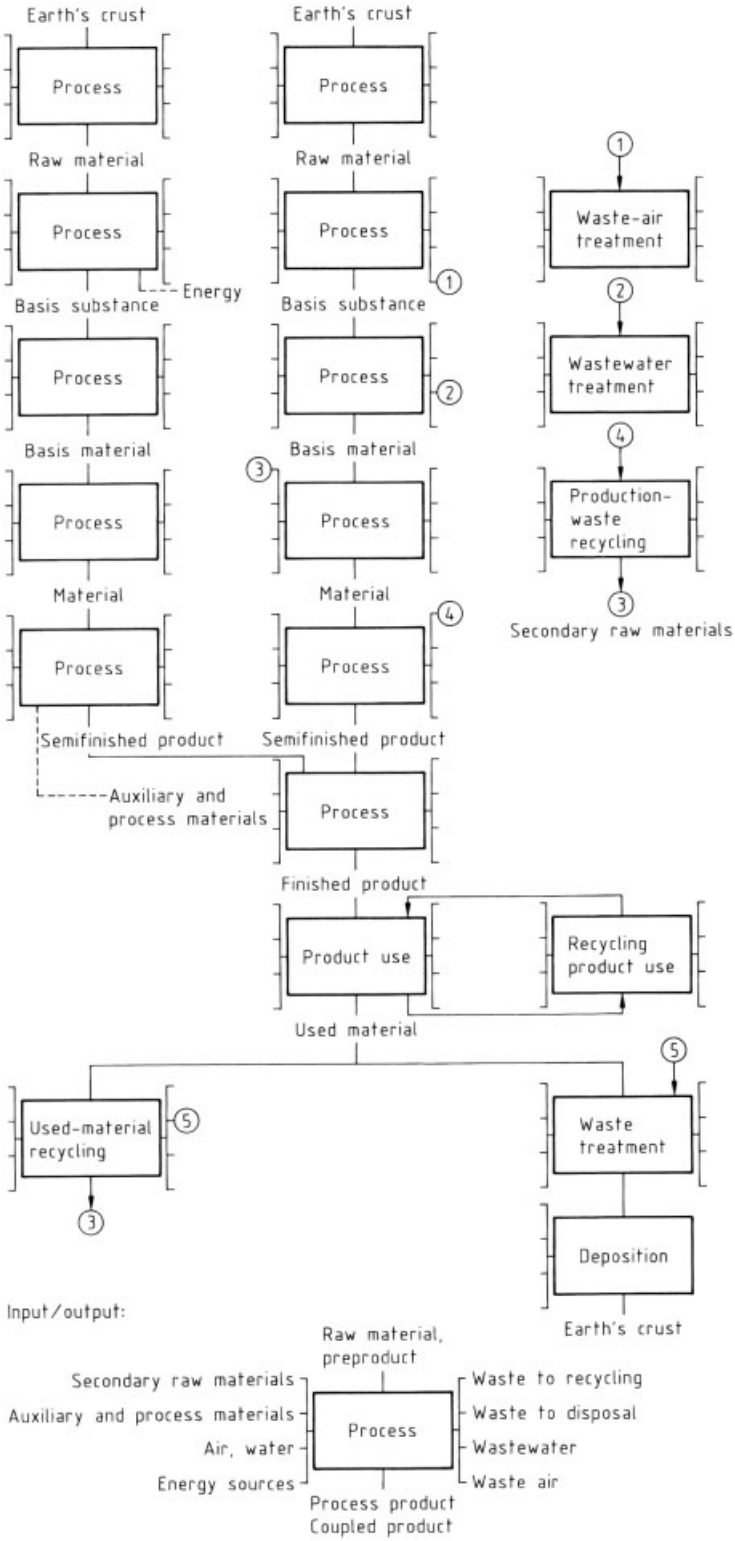

Figure 7. Schematic diagram of a greatly simplified product line with basic structure and examples of side structure (energy, auxiliaries, and media 1 to 5) [50]

4.3.2.3.1. Process Description

The description of a process includes the following elements:

1) Site (geographic system boundary)
2) Components involved (as a rule these are components of the plant) and interfaces between the process and upstream and downstream operations (process system boundary)
3) Nature of the process (production, consumption, or disposal process; involving transformation of substances, forming, transportation, etc.), operations involved in the process, and process parameters (pressure, temperature, etc.)
4) Nature of substances and energies used; results of the process
5) Disturbances, i.e., departures from normal operation, with their duration and frequency
6) Source information

For any specific LCA, the concrete process is described. For generalized LCAs, the process description can be extracted from the literature (as it can in specific LCAs when processes are not accessible). A problem arises when average values are cited for different process sequences that lead to the same process result but cannot be traced back to individual processes. Such data may include material and energy information that does not even apply to one of the process sequences included in the average. The process description includes initial data such as the site, but it also forms the basis for further identifications, for distribution functions (Section 4.3.2.4.2), and also for the actual calculations. An adequate process description is the first step in making the inventory analysis reproducible [31], [41].

4.3.2.3.2. Acquisition of Basic Data

A precondition for the acquisition of specific basic data [41] is the early motivation of the specialists needed for data preparation who work on the processes being assessed in the companies concerned. Data acquisition includes material and substance data as well as energy data. Here "material" refers to identifiable, delimitable, and assessable matter that may consist of one or more than one substances. In this sense, a "substance" is a chemical compound or a chemical element.

The term "basic data" refers to data as acquired from companies and not transformed in any way. Basic data in original form, together with units and reference quantities, are entered in the "basic data" process table and form the basis for the inventory analysis. Some basic data can be inferred from the process description (Section 4.3.2.3.1) and are acquired in the businesses concerned when the specific inventory analysis is done. Ranges of variation of process data (inexact data) must be considered along with the frequency and duration of scheduled disturbance. Here "disturbance" denotes departures from steady-state operation, such as startup and shutdown phases. Departures occasioned, e.g., once or twice a year by readjustment of thermal production plant must be known and taken into account. Disturbances that occur more often can be compensated if data acquisition takes place over a sufficiently long time period so that the resulting errors become negligible. When no operating data can be acquired, literature data and third-party information are used as the data basis. Such data must be examined for plausibility and identified as specific or generalizing data. The age and the actual origin of these data must be considered and must be stated in every instance. When anonymized data are employed, this requirement cannot be met [31], [40].

When data varying in quality are used, or when the level of detail varies, the data must be made symmetrical; otherwise, the comparison may indicate that the worse process or even the worse total product is precisely the one for which the data depth is greatest [39]. The handling of data gaps is discussed in Section 4.3.2.4.2.

4.3.2.4. Material and Energy Balance

4.3.2.4.1. Process Balance

A "balance" is a comparison of the inputs and outputs of a process. The balance can be prepared for normal operation and for process disturbances, such as departures from normal operation. Process inputs and outputs are materials, substances, energies, and information. In the framework of the inventory analysis, a mass balance is prepared and then forms the basis for

an energy balance. The mass (material) balance is carried out on the basis of materials. A material can consist of one or more substances, a substance being a chemical compound or an element. Input materials for a mass balance include raw materials and products from upstream operations as well as secondary raw materials, auxiliaries, media, air, water, and energy resources. Output materials include the process product or coproducts, which may also be feeds to a downstream process, as well as wastes for recycling, wastes for disposal, wastewater, and waste air (Fig. 7). The forms of energy to be balanced include chemical energy, thermal and electrical energy, radiation, and noise.

The production and disposal of equipment, unless otherwise stated in the goal definition, are not considered in the LCA. Equipment itself may, however, be the object of investigation for an LCA. Here "equipment" includes machinery, buildings, tools, and other technical plant.

The space requirement, e.g., for original production (mining, agriculture, forestry, crude oil production, extraction of other raw materials, etc.) and the area used for disposal must be quantified as far as possible for the impact assessment phase, and the degree of "naturalness" must be identified ("hemerobic" levels [32]).

The material and energy balances are based on two fundamental laws, those of the conservation of mass and of energy. For each process, these imply the balance condition that the sum of the input masses m_I must be equal to the sum of the output masses m_O plus the accumulated masses m_A (Eq. 1). If there is no accumulation during the balance interval, the balance equation simplifies accordingly. Accumulation can be neglected if the balance interval includes a sufficiently large number of resettings of the plant [31], [41]. This balance condition is the prerequisite for the clarity and reproducibility of the life-cycle assessment.

$$\sum m_I = \sum m_O + \sum m_A \tag{1}$$

For the energy balance the equality of the total inputs and the total outputs is dispensed with, but not the plausibility check, because the energies are allocated as far as possible to the energy resources and discrepancies in the balance are made up by heat losses. The steps in process balancing are as follows:

1) Prepare the basic data list for materials, substances, and forms of energy (Section 4.3.2.3.2)
2) Establish the nature (process goal), mass, and unit of the process reference quantity
3) Conversion of basic data and process data to consistent mass or energy units and a consistent process reference quantity (process goal)
4) Process modeling and data correction, including the closure of data gaps [31], [41]

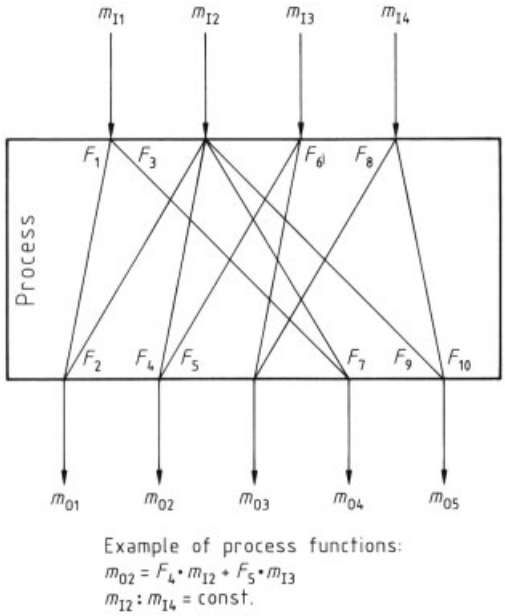

Figure 8. Principle of process modeling with process functions (distribution function and input ratio) [41]

4.3.2.4.2. Process Modeling and Data Correction

The basic data set generally does not satisfy the balance condition (Eq. 1). Data correction and the closure of data gaps are accomplished with distribution functions, which are written in the form of a consistent system of equations for each process. The system of equations is normalized to the process reference quantity (e.g., 1000 kg of process product). The equations are linear approximations to the actual process sequence

within certain constraints that represent the dependences of the output material streams on the input material streams as well as those between input material streams (Fig. 8). In order to write the system of equations, the process in question must be known as well as possible. The equations are based on the real process steps, e.g., on chemical reactions, and on material distributions occurring in practice. They cannot, however, reflect the actual process steps as they are influenced by various process parameters, such as pressure and temperature. A process simulation is too expensive for the majority of processes and would not adequately improve the result of the inventory analysis [41].

As the balance condition is the basis for the clarity and reproducibility of the LCA, process modeling cannot be dispensed with if the model of the inventory analysis is to be used for optimization calculations and sensitivity analyses.

For inexact data, such as averages with their associated ranges of validity, no computational rules have yet been devised for integrating this type of data in the inventory analysis. Doing so would again yield results in the form of ranges. Today, all that can be done is to calculate with averages or with best-case or worst-case alternatives. Each of these three approaches fails to reflect reality, since there is no product line in which just one of the assumptions applies to evey process. It must be supposed that computational rules will soon be developed such that inexact (fuzzy) data can be used in inventory analyses [39].

4.3.2.5. Allocations

In the data acquisition process, balance data are first related to the respective processes (process allocation), then to the object of investigation (system allocation). On the basis of the functional unit, the unambiguous balance boundaries for the system are determined, including their extensions for complementary goods. The functional unit, as a type of normalization basis, ensures that the benefits due to the several systems are equal, which is the prerequisite for comparability (Section 4.3.2.5.4). No balance data are allocated to the functional unit.

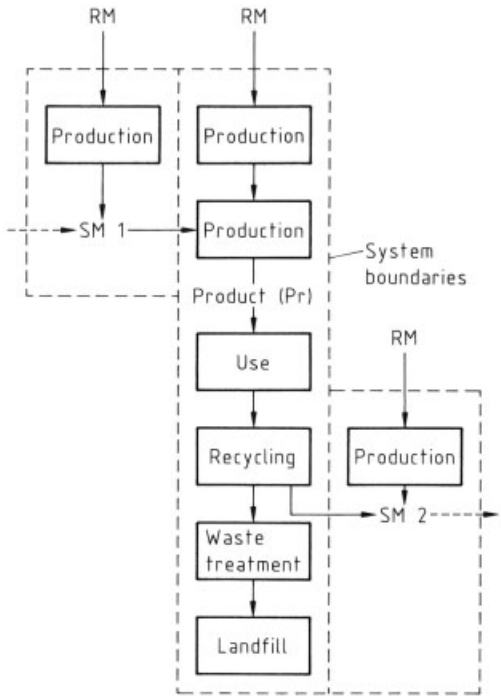

Figure 9. Schematic product line for product Pr with inlet and outlet secondary raw materials (SM 1, SM 2)

4.3.2.5.1. Process Product and Coproducts

Process allocation is performed for processes whose material flow is directed toward the object of investigation (production processes) – the process product. For processes whose main material stream is directed away from the finished product (post-consumption processes and disposal processes for production), allocation is to the input to be reclaimed or treated [41].

When processes generate more than one process product (i.e., coproducts) or take as inputs more than one upstream product to be recycled or treated, there is more than one process reference. The allocation of material and energy streams and thus of environmental burdens up to this process must be split among all the coproducts. The share of process product x (Sx) should preferably be in proportion to the masses of process products of the respective process (Eq. 2). This splitting method, which is crucial for the result of the inventory analysis and thus for the LCA, cannot be done according to masses for every process that yields coproducts. Splitting

by volume, energy content, and possibly other physical quantities is also possible [43]. The selection of the governing quantity must be done in a justifiable and traceable way. For processes where the process reference is on the input side and there is more than one upstream product, splitting is done similarly to the coproduct case.

$$S_x = \frac{m_x}{\sum_{\forall i} m_i} \tag{2}$$

where m_i is the coproduct of the process under consideration.

4.3.2.5.2. Object of Investigation

The following method is used to allocate balance data to the object of investigation. For each pre-consumption process, the mass of the pre-product for process $i-1$ (counting from the object) is divided by the mass of the product from process i. Qualitatively, the process product of process i becomes the pre-product of process $i-1$. Starting from the object of investigation, the "quantity multiplier" QM for the n-th process is the product of all the quotients from the pre-consumption processes, from manufacture of the object of investigation up to the n-th process (Eq. 3). All data for this n-th process are multiplied by the previously determined multiplier to determine the contribution of this process relative to the object of investigation. In the case of processes whose material flow is away from the object of investigation, the quotients for determining the quantity multiplier are formed in a different way: the mass of the process product of the preceding processes $i-1$ is divided by the corresponding input mass of process i (Eq. 4). The process goal, in contrast to production processes, is directed toward the pre-product. Qualitatively, the process product of process $i-1$ becomes the pre-product of process i. Processes of additive environmental protection in the area of preconsumption, or processes for generating energy and producing auxiliaries and media in the area of post-consumption, have quantity multipliers that are calculated with appropriate mixed forms of Equations (3) and (4) [41].

$$\mathrm{QM}_{n,\,\mathrm{pre-consumption}} = \prod_{i=1}^{n} \frac{m_{\text{pre-product of the following process } i-1}}{m_{\text{process product of process } i}} \tag{3}$$

$$\mathrm{QM}_{n,\,\mathrm{post-consumption}} = \prod_{i=1}^{n} \frac{m_{\text{process product of the following process } i-1}}{m_{\text{pre-product of process } i}} \tag{4}$$

In this way the material and energy data of all processes within the product line are allocated to the object of investigation. The further environmental burdens resulting from the production of complementary goods to guarantee equality of benefit (Section 4.3.2.5.4) are determined by the same rules and added to the environmental burdens of the product line of the object of investigation. This is how, in the framework of the functional unit, with the exception of the case of open-loop recycling (Section 4.3.2.5.3), the environmental burdens are determined for which the impact potentials are found in the classification and characterization steps.

4.3.2.5.3. Open-Loop Recycling

Open-loop recycling is characterized by a secondary material leaving the balance system or a secondary material entering the balance system, and similarly for energies. Figure 9 shows a product line with the product Pr, the primary benefit of the system, taken as the object of investigation. After consumption, a secondary raw material SM_2 is generated by recycling associated with the product line. This secondary material is reused not in the same product line but in another, unknown, product line. For consistency in the handling of this additional benefit (secondary benefit) supplied by and leaving the system under consideration, the material and energy streams and thus the environmental burdens found in the LCA are those that would occur if this secondary material were produced from primary raw materials (Eq. 5). To this end, the masses m of primary raw materials (RM) as well as wastes and emissions (WE) arising in the respective time intervals are used, namely $(t_p - t_0)$ from extraction of raw materials to production and $(t_\infty - t_0)$. As shown in Equation (6), the sum of the environmental burdens (EB) so produced but not yet weighted with the impact potentials (see Section 4.3.3) are subtracted from the environmental burdens of the product line under

consideration (m_{Pr}) [44]. Similar formulas apply to energies.

$$m_{\mathrm{SM}} = \sum_{t_0}^{t_{\mathrm{P}}} \sum_{i} (m_{\mathrm{RM}} + m_{\mathrm{WE}})_{t,i} \tag{5}$$

$$EB_2 = \sum_{t_0}^{t_\infty} m_{\mathrm{Pr}} - \sum_{t_0}^{t_{\mathrm{P}}} m_{\mathrm{SM}_2} \tag{6}$$

Also shown in Figure 9 is a secondary raw material SM_1 entering the balance system. It is not known what environmental burdens arose in the production of this secondary raw material. For consistency in balancing, the environmental impacts are determined for the case that this secondary raw material is produced from primary raw materials. The environmental burdens so determined for the secondary raw material are added to the environmental burdens of the product line (Eq. 7). For the case where the secondary raw material SM_2 leaving the system is equal in quality and quantity to the secondary raw material SM_1 entering the balance system, the pluses and minuses in the environmental burden calculation would offset one another. This represents the ideal case of closed-loop recycling [44]

$$EB_1 = \sum_{t_0}^{t_\infty} m_{\mathrm{Pr}} + \sum_{t_0}^{t_{\mathrm{P}}} m_{\mathrm{SM}_1} \tag{7}$$

Only when the secondary raw material replaces a material can one speak of recycling, and only then can these allocation rules be applied in a methodologically correct manner.

4.3.2.5.4. Equality of Benefits

The functional unit is intended to ensure equality of benefit between the systems under consideration [42], [45]. This happens in the product life-cycle assessment based on the product used for comparison. Systems being compared must be equal in benefit; this means, among other things, that products being compared must meet the same requirement profile to have equal primary benefits. Equality of benefit exists in relation to the objects of investigation provided there is no additional benefit (secondary benefit).

Additional benefit is created, in particular, by recycling processes, but also during the primary production of byproducts (e.g., steam) that cannot be allocated by simple coproduct relations (Section 4.3.2.5.1). In the case of open-loop recycling (Section 4.3.2.5.3), e.g., an additional benefit is created by the secondary raw material SM_2 (Fig. 9), and this additional benefit is not taken into account by the product as reference quantity. Equality of benefits between systems exists when the sum of goods produced in each system is equal in quantity and quality. For this to be so, the system must be expanded to include the processes for producing the so-called complementary goods C of equal quantity and quality from primary raw materials (Fig. 10). Only by this procedure are systems that have the same primary benefit but are not equal in benefit because of the secondary products produced (Eq. 8) rendered comparable (Eq. 9) and thus satisfy the functional unit.

$$\mathrm{A}_1 \cup \mathrm{B}_1 \neq A_2 \tag{8}$$

$$\mathrm{A}_1 \cup \mathrm{B}_2 = \mathrm{A}_2 \cup \mathrm{B}_\mathrm{C} \tag{9}$$

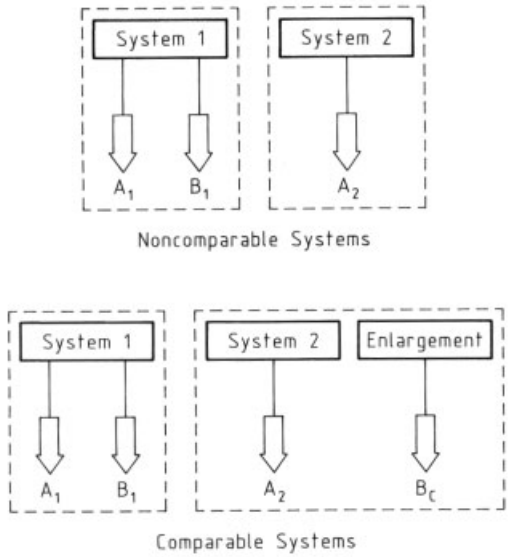

Figure 10. Expansion of system to render noncomparable systems comparable [42]

What is compared is not the benefits of the systems – these are equal – but the environmental burdens generated by systems of equal benefit.

Table 10. Some indicators [30], [32], [47], [51–53]

Heading	Indicator (reference substance)	Spatial relevance
Resource consumption		global
Renewable resources	area required, biotic depletion factor	
Nonrenewable resources	range, abiotic depletion potential	
Greenhouse effect	global warming potential (CO_2)	global
Ozone depletion	ozone depletion potential (R-11)	global
Acidification	acidification potential (SO_2)	continental/regional
Eutrophication	nutrification potential	continental/regional
Terrestrial	(nitrogen)	
Aquatic	(phosphate or COD)	
Photochemical smog	photochemical ozone creation potential	
	(ethylene)	regional
Human toxicity	human toxicity potential	continental/regional
Ecotoxicity		continental/regional
Terrestrial	terrestrial ecotoxicity potential	
Aquatic	aquatic ecotoxicity potential	
Area used/species diversity	m^2 (hemerobic levels)	regional/local
Odor	odor threshold value	local
Noise	acoustic level in $Pa^2 \cdot s$	local

4.3.3. Impact Assessment

In the impact assessment step, potential environmental burdens of the quantities of materials and energies determined in the inventory analysis are aggregated on the basis of scientific analyses of environmental processes [46]. For practicability and feasibility, materials and substances are rendered comparable in potential impact by being allocated to a category and thus given the same unit so that they can be aggregated. Actual impacts on the environment due to the systems (product lines) being compared are not determined. The purpose in determining the impact potentials is to eliminate subjective effects in classification and characterization. According to SETAC (Society of Environmental Toxicology and Chemistry), these two operations are described as follows [28]:

1) Classification: selection of the impact categories (headings) for examination and allocation of inventory analysis data to a comprehensible number of headings (Table 10)
2) Characterization: definition of scientific units of measure or indicators for each heading

With the indicators chosen, the inputs and outputs are then recalculated as impact potential quantities, which are aggregated in the " ecoprofile of products."

4.3.3.1. Classification

In classification, data from the inventory analysis step are allocated to "headings" [28] and the decision is made as to what headings will be used in the life-cycle assessment [47]. The allocation of cumulative resource consumptions and emissions over the life cycle to concrete environmental problems is intended to give a more direct relationship to environmental problems and policies without restriction to individual media (air, water, soil). This approach means that aggregation can be done on a scientific basis, because the action chains for individual materials exhibit similarities that can be described in a scientific manner.

In the LCA context, various headings differing in spatial relevance are considered (see Table 10). The spatial relevance is crucial because, in concrete ecosystems, the prestress, the medial propagation capabilities, and the buffer capacities can vary widely from one pollutant to another. For this reason, the use of system inventory analysis data is methodologically dubious, particularly in the case of local and regional headings, when the spatial relevance of the inventory analysis data (geographical balance space) and the spatial relevance of the headings (Table 10) are different.

The selection of headings for consideration depends on the objective of the LCA and takes account of production-specific burdens.

4.3.3.2. Characterization

Global impact indicators for the greenhouse effect and ozone-layer damage have been developed by, respectively, the Intergovernmental Panel on Climate Change (IPCC) and the Scientific Assessment Panel under the tutelage of the World Meteorological Organization [47]. These

indicators have been acknowledged internationally.

Several impact indicators have the feature that relative impact potentials are stated in comparison with a reference substance. This is especially so for LCAs covering large geographic areas (country, continent). Here the ecosystems, organisms, and receptors involved and the conceivable propagation paths for substances (given their different conversion reactions and action mechanisms) are so diverse that the concrete impact can no longer be described. What is more, the specific impact cannot be the goal of an LCA; this is a contrast between LCA (lessburdens-are-better approach) and environmental compatibility testing (only-burdens-above-threshold-are-considered approach). The extension to actual impacts is not possible with the existing indicators. Uncertainties and limitations of the environmental indicators must be considered in their interpretation.

The indicators of the headings should be free of normative and political considerations. The following parameters are to be taken into account [48]:

1) Harm (to human beings, fauna, flora, buildings, the ozone layer, etc.)
2) Persistence in the environment
3) Mobility (danger of long-lived emissions becoming spread over large areas)
4) Accumulation in the environment
5) Synergisms/antagonisms or reactions with other substances (e.g., formation of photochemical smog)

Action chains and feedback mechanisms must also be examined whenever possible (Fig. 11).

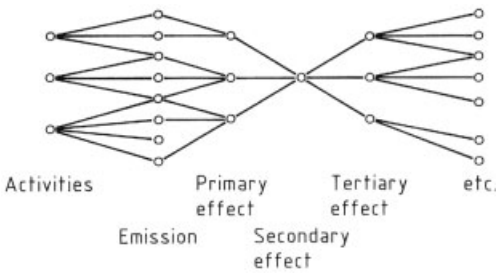

Figure 11. Example of an action chain [49]

For example, the action chain for the greenhouse effect can be described as follows [50]:

– Various activities, such as burning of fossil fuels, lead to
– Emission of gases, which have various
– Primary effects. They can directly or indirectly absorb IR radiation and thus upset the equilibrium between energy absorption and energy emission.
– The secondary effect is a rise in the mean temperature of the Earth.
– Tertiary effects exhibit geographic dependences: melting of ice, rise in sea level, drought, etc. Further consequential impacts include changes in the composition of vegetation.

This example makes it clear that the impacts at the beginning of the action chain are chemical or physical in nature, while the derivative impacts are biological and may be functions of geographic location. Quantification becomes more difficult down the action chain, so that the definition and precise demarcation of the heading is important. The indicator should bear on a "node" in the action chain (on the secondary effect in the above example). Table 10 summarizes some of the indicators developed to date and shows the spatial relevances of the headings.

The reductionist approach of [54] restricts itself to a few parameters for simplicity's sake: consumption of renewable and nonrenewable energy, consumption of minerals and water, wastes (classified as municipal, hazardous, and radioactive), global warming potential, ozone depletion potential, acidification, and eutrophication. The last four of these indicators are described below.

4.3.3.2.1. Greenhouse Effect

The heating m_j of atmospheric layers near the surface of the Earth when long-wavelength radiation emitted from the surface is absorbed by greenhouse gases is described by the global warming potential (M_{GWP}) of all greenhouse substances [30]:

$$M_{\mathrm{GWP}} = \sum_{\forall j} m_j \cdot \mathrm{GWP}_j \tag{10}$$

The direct global warming potentials GWP_j of various infrared-absorbing substances are stated

relative to CO_2 for an integration time of 100 years:

$$\mathrm{GWP}_j = \frac{\int_0^{100} a_j c_j \mathrm{d}t}{\int_0^{100} a_{\mathrm{CO_2}} c_{\mathrm{CO_2}} \mathrm{d}t} \tag{11}$$

where a_j is the absorbing power and c_j the concentration of substance j at time t.

4.3.3.2.2. Ozone Depletion

The decline in the stratospheric ozone concentration is described by the ozone depletion potential (ODP), which is the ozone destruction efficiency relative to R-11 ($CFCl_3$). This value is determined in one- or two-dimensional computer simulations [49], which are necessary because the ODP is governed by physical and chemical processes in the atmosphere. An upper boundary for the ODP can be calculated from the relative quantity of chlorine that can reach the stratosphere as a result of emission of a gas (chlorine loading potential CLP, also relative to R-11).

The Swedish Environmental Research Institute has proposed that ozone depletion should be described in terms of three separate figures – an indicator for substances containing chlorine (ODP, CLP), one for those containing bromine (ODP), and the total quantity of N_2O – until further research makes it possible to combine all this information into one parameter [49].

4.3.3.2.3. Acidification

If the acid content of soils and waters rises excessively, mineral nutrients may be leached from soils and heavy metals may be mobilized. The reference for an impact indicator of acidification (M_{AP}, where AP stands for acidification potential) is the potential of SO_2 to form H_3O^+ ions. The acid-forming pollutants included in the inventory analysis are aggregated as [30]

$$M_{\mathrm{AP}} = \sum_{\forall j} m_j \cdot \mathrm{AP}_j \tag{12}$$

This formula is based on complete oxidation of sulfur dioxide in air followed by the reaction with water (raindrops, atmospheric water vapor), yielding 2 mol of H_3O^+ per mole of sulfur dioxide [30]. To allow for the difference in activity between substances j, it must be taken into account that the loss of a hydrogen ion is an equilibrium reaction described by a substance-specific acidity constant K_{S}, and that the fraction of H_3O^+ ions liberated is a function of pH [55]. The relative acidification potential is then found by dividing n_j, the number of moles of H_3O^+ liberated per mole of substance j, by the number $n_{\mathrm{SO^2}}$ of moles of H_3O^+ per mole of SO_2:

$$\mathrm{AP}_j = \frac{n_j}{n_{\mathrm{SO_2}}} \cdot \frac{64}{M_j} \tag{13}$$

4.3.3.2.4. Eutrophication

If additional nutrients are introduced into soils and waters, excessive growth of certain species (e.g., algae) can disturb the ecological equilibrium. The action of eutrophying substances M_{eu} is determined by aggregating as follows [30]:

$$M_{\mathrm{eu}} = \sum_{\forall j} m_j \cdot \mathrm{NP}_j \tag{14}$$

The eutrophication of substances (where NP stands for nutrification potential) shows a strong receptor dependence. In terrestrial ecosystems, the deficiency of nitrogen for eutrophication is dominant [49], so that nitrogen emitted into the air must be converted to nitrogen equivalents (molecular mass of nitrogen = 14):

$$\mathrm{NP}_{\mathrm{terr},j} = \frac{n_j \cdot 14}{M_j} \tag{15}$$

where n_j is the number of nitrogen molecules in substance j.

In aquatic ecosystems, the controlling parameter for eutrophication is oxygen depletion. In addition to direct oxygen consumption in the degradation of organic materials, a second problem is the indirect oxygen demand for the degradation of biomass built up via the emitted nitrogen (deficiency factor in marine ecosystems) and phosphorus (deficiency factor in limnetic ecosystems). For the biomass generated, it is assumed that the formation of one mole of biomass requires 16 mol of nitrogen and 1 mol of phosphorus [49].

4.3.4. Interpretation

The steps of classification and characterization help to reduce the number of substances and energies identified in the inventory analysis to an easily handled number of headings for each product investigated. If there is no one product that has better values than another in all the headings, determination of the ecologically "better" variant requires an interpretation so that the results found in the various categories can be combined.

4.3.4.1. Objective and Requirements

The purpose of the interpretation step is to apply objective scales to arrive at a relative weighting of the results obtained in classification and characterization, as well as inventory analysis results that cannot be put in the form of action potentials [27]. The DIN standard points out that an interpretation justified in purely scientific terms is not possible and any valuation must have a political aspect. This means that the interpretation must integrate scientific scales with political environmental quality goals. These goals or targets are embodied in international conventions (Agenda 21, the Rio Declaration, the biodiversity treaty [56], [57], etc.), human rights, principles of environmental protection (precautionary principle), and national initiatives.

The interpretation step is intended to provide further aggregation of the data so as to reduce the complexity of environmental burdens and thus serve as an aid to decision-makers. Interpretation does not constitute a decision.

Interpretation makes it possible to arrive at subjective assessments in a context-independent, reproducible, scientifically sound way, as well as to generate subjective, context-dependent assessments. The methods of interpretation must satisfy strict requirements as to clarity, reproducibility, plausibility, and discussability. Subjectivity in interpretations does not mean arbitrariness; it rather expresses an awareness that scientific knowledge of the complex relationships prevailing in the environment is limited.

4.3.4.2. Interpretation Approaches

There is as yet no consensus on interpretation methods [29]. Three main lines are recognized:

1) Normalizing, quantifying assessment
2) Verbal-argumentative assessment
3) The reductionist approach

Mixed forms of normalizing-quantifying and verbal-argumentative assessment are applied where the aspects that cannot yet be quantified are interpreted by verbal-argumentative techniques.

4.3.4.2.1. Normalizing, Quantifying Methods

The assessment of environmental factors by quantitative normalizing methods is carried out by means of a model and by using limit values [58], quality targets [59], degrees of goal fulfillment (utility analysis) [60], critical loads [55], [61], loss – benefit analysis [62], monetarization [63], and so forth. These methods can be classified as follows:

1) Single-index methods [29] that yield, at the end of the valuation, a number representing the (possibly relative) environmental burdens of the items compared
2) Partial aggregation methods in which the environmental burden is expressed in terms of several parameters
3) Methods regarded as single-index or partial aggregation techniques but omitting the classification and characterization steps

Indicators derived in classification and characterization can be directly weighted and aggregated into a single index by using utility analysis with quantitative degrees of goal fulfillment [60] or by attempting a quantitative operationalization of the notion of "sustainable development" [55]. To date, however, normalizing, quantifying interpretation methods have mostly employed one of the two BUWAL concepts: aggregation to four parameters (critical quantities of air and water, specific landfill volume, energy equivalence [58]) or to one parameter (" ecopoint" method [59]); this has been used most often in LCAs not requiring a classification and characterization in the sense of SETAC [28]. The critical

volumes are the theoretical quantities of air and water that would be required to reach the politically specified limit for substance *i* by dilution (Eq. 16). These quantities do not represent impact potentials but must be understood as pure computational values for orientation. The energy equivalence value references all forms of energy to the primary energy used. The "ecopoint" method based on this applies " ecofactors" (Eq. 17) to reduce the environmental burden to a single parameter (Eq. 18). The "critical flow" in the denominator is the politically specified maximum permissible burden for the assessment space (e.g., Switzerland) and does not represent the real impact potential. Alternatively, the scientifically derived "critical loads" [61] could be used to determine the ecological shortage [55]. Critical loads are the estimated pollutant exposures, derived from scientific models, such that no perceptible changes in ecosystem structure and function are anticipated according to present knowledge. Among the objects for comparison, the one with the lowest ecopoint number is deemed the most "environmentally friendly."

$$\text{Critical volume} = \sum_{\forall i} \frac{\text{Emission}_i}{(\text{limit value})_i} \tag{16}$$

$$\text{Ecofactor} = \frac{\text{present pollutant flow in assessment space}}{\text{critical flow}} \cdot \frac{\text{dimensional factor}}{\text{critical flow}} \tag{17}$$

$$\text{Ecopoint} = \sum_{\forall j} (\text{Ecofactor} \cdot \text{Emission})_j \tag{18}$$

4.3.4.2.2. Verbal-Argumentative Assessment

In verbal-argumentative methods [64], predefined goals and interpretation criteria are applied in an attempt to assess the relative ecological weights of the headings on the basis of well-defined environmental goals. To ensure the reproducibility of verbal-argumentative methods, a qualitative (i.e., described merely in verbal form) utility analysis is employed as a structuring aid.

The starting point for a verbal-argumentative assessment is often a normalization, which may be done either with the "fir tree" representation [65] or by normalization [30], [66] of the impact indicators. The tree diagram is based on the magnitude ratio of the action indicators of the comparison object. A branch of the tree is plotted only for the comparison object that shows greater burdens under this category. In normalization, the magnitudes of the impact indicators are referred to the total for this heading for a certain region (global [66] or national [67]) and a certain time interval (year). Normalization can also be performed as part of the classification and characterization steps.

On this basis, the interpretation can be performed outside or inside the area of responsibility of the party carrying out the life-cycle assessment. Techniques available are Delphi, expert hearings, ranking, and so forth [68–70]. The most advanced of these are the UBA (German Federal Environmental Agency) method [65] and analogous procedures for European "ecolabeling" (without classification and characterization) [71]. The first step in these methods, weighting, is done similarly to normalization: in terms of the specific contribution or the production intensity (PI) factor. These quantities reflect the difference from one product group or sector to another in relevance to the corresponding national or European total emission. Next, the ecological significances of the headings or substances are determined, discussed, and evaluated on the basis of definite criteria. The UBA method uses the following criteria: ecological hazard potential, reversibility, spatial relevance, environmental preferences of the public, and relationship of the prestress to quality goals. These two factors (specific contribution versus PI factor and significance factor) are combined in a verbal-argumentative manner to reflect the weighting of the categories. Now, for each comparison object, it is determined how much of the relative disadvantage in an impact category or heading (symbolized by the presence of a branch in the fir tree model) can be offset or even priorized by the relative advantage in another category. The procedure for deriving the ecological significance is under scientific discussion and a topic of ISO and DIN activity.

The verbal-argumentative interpretation method expresses weights in the form of words (slight, moderate, great significance, etc. [65]), symbols (+, −) [72], letters (ABC weak-point

analysis, where A = critical, B = may be critical, C = noncritical) [33], or numbers.

4.3.4.2.3. Reductionistic Approach

Because there is no consensus on interpretation, this step is sometimes omitted altogether (along with classification and characterization) [73], or only a subset of the inventory analysis is considered, such as resource consumption [74] or ecotoxicity [75]. The reasons are that although quantifying-normative methods do offer the advantages of systematic character and reproducibility along with a transparent and easily comprehended representation of results, at the same time they have the disadvantage of an illusory reliability and the apparent unambiguity of the results, especially with single-index methods. Verbal-argumentative methods lack these disadvantages but are not reproducible because of their subjective elements.

If interpretation is dispensed with, at least a weak-point analysis can be carried out to identify "ecological hot spots." Sensitivity analyses can aid the evaluation of optimization proposals. Inventory analysis and impact assessment by themselves will be important aids to decisionmakers [31].

The reduction of the ecological analysis to resource consumption is justified by the essential impossibility of "scientifically studying . . . or quantifying . . . all chemical, physical, and biological impacts on ecosystems, even for a single substance released into the environment." Instead, the material or area intensity per service unit is used as the sole decision parameter. In this treatment, all materials and areas used are added together in an undifferentiated manner, i.e., without classification and characterization and without regard to present knowledge of toxicity and other environmentally relevant aspects [74].

4.3.5. Improvement Assessment (Optional)

According to SETAC [28], it is only in the improvement assessment phase that options for reducing environmental impacts are identified and evaluated. The study of ecological optimization can be done by comparing several options (e.g., alternative processes) or performing a sensitivity analysis if the functional relationships have been modeled (see Section 4.3.2.4.2). According to DIN [27], the optimization step can also be carried out as part of the inventory analysis if what is desired is a reduction in, e.g., the consumption of special raw materials or the emissions of certain substances. Ecological optimization, according to DIN [27], is done in the context of interpretation.

4.4. Critical Review

To determine whether the specifications set down in goal definition and scoping have been met, the method has been properly applied, and the data used are plausible, a critical review is carried out. The selection of the examiner is important for plausibility checking of LCA results. For life-cycle assessments within the company, an employee can perform the review. In the case of public LCAs, the examiners must be selected from among interested circles. The review can take place in parallel with the LCA or after it has been completed. To save time and avoid duplication, parallel review accompanying LCA performance is preferable to post-assessment review.

The parallel critical review takes place in the following phases of the life-cycle assessment:

1) At the beginning, to verify the selected functional unit and the system boundaries as well as the planned volume of data and the desired data quality. In public LCAs, an additional purpose is to involve the technically qualified public early (including circles with ecological, social, and economic interests) and promote acceptance [27].
2) After the first data collection and modeling, to check the specific method used and also to pick up hints for subsequent processing.
3) At the final report stage, to check the whole LCA (including conclusions) for plausibility [28].

5. Waste Management in the Chemical Industry

5.1. Introduction

The central function of the chemical industry is the economical transformation of substances into products desired in the market. This transformation should take place with a yield as close to 100 % as possible, but there is always a percentage of undesired products, which are collected under the term "residues."

A second important task of the chemical industry is to minimize the impacts of residues on the environment as far as is technically feasible. To carry out this task, the industry employs waste-management concepts which encompass waste avoidance by avoiding or reducing the quantity of residues and the recycling of residues, as well as the disposal of the remaining residues as "waste." Avoidance, reduction, and reclamation take priority over disposal by incineration and landfilling. A final task of waste management is avoiding or reducing the quantity of production wastes and disposing of them safely.

In the downstream processing, conversion, and consumption sectors, some product becomes waste, and this is referred to as "product wastes" to avoid confusion with production wastes. Recycling of product wastes is acquiring greater importance and is becoming a further waste-management function in the chemical industry.

5.2. Chemical Industry Wastes

The industrial production sector in Germany generated 259.1×10^6 t of waste in 1990 [76]. This included 8.95×10^6 t of waste requiring special monitoring, which is also referred to as hazardous waste. In the same period, the German chemical industry, as part of the production sector, generated 8.06×10^6 t of waste [76], including 2.7×10^6 t of hazardous waste. The broad production spectrum of the industry means that the wastes are also very diverse in terms of chemical composition and consistency. This is illustrated by the following list of the most important wastes (1990, in 10^3 t) [77]:

Sewage sludges, sludges from wastewater and water treatment (dry basis)	1174
Building rubble, excavated soil, pavement rubble without harmful contaminants	1159
Wastes from energy production (slags, ashes, filter dusts)	883
Red mud	759
Residential-type commercial waste (incl. used paper, packaging, industrial sweepings)	522
Building rubble, excavated soil, soils with harmful contaminants	331
Gypsum wastes with harmful contaminants	312
Inorganic acids, acid mixtures	278
Gypsum/lime mud with harmful contaminants	235
Sludges from precipitation and solution processes with contaminants	201
Organic solvents and solvent mixtures containing halogenated solvents	131
not containing halogenated solvents	145
Filter aids and absorbents with contaminants (kieselguhr, activated carbon, activated earths)	144
Organic distillation residues containing solvents	128
Carbide sludge with no harmful contaminants	73
Barium sulfate sludge and rock-salt residues (gangue)	63
Plastics wastes	61
Packaging material with harmful contaminants	40
Wastes from plastics manufacturing and converting	32
Lime and gypsum mud with no harmful contaminants	22
Oil and gasoline separator contents	14
Organic dyes	12
Plastics slurries	12
Aliphatic amines	11
Filter dusts containing nonferrous metals	11
Sludge from tank cleaning	9
Other wastes	1298
Total	*8060*

Over half the quantity of wastes is accounted for by three large groups: 29 % wastes not associated with chemical production (building rubble, excavated soil, wastes from energy production facilities); 15 % from wastewater and water treatment; and 15 % inorganic sludges.

This quantitative information represents a waste balance which deviates from the waste statistics, and is defined as follows [76]:

1) Duplication is removed by subtracting wastes accepted by other businesses, and secondary wastes generated in waste treatment (e.g., slag and ash from incineration)
2) Quantities delivered by waste producers to reprocessors or scrap dealers (i.e., residues) are not included
3) Sewage and water-treatment sludges are reported on a dry basis because of their high water content

Comparable statistics broken down in a similar way are not available for the chemical industries of other countries.

5.3. Waste-Management Concepts

5.3.1. Residues and Wastes from Production

A chemical production process generates residues in addition to the intended end product. These residues can be recycled (Fig. 12). If recycling is impossible for technical or economic reasons, residues become waste. In addition, polluted wastewater and waste gases can be formed. Disposal of wastes includes incineration and landfill disposal. Waste incineration generates secondary wastes in the form of slag and ashes, which have to be utilized or disposed of. Wastewater is produced in waste-gas scrubbing. The treatment of wastewater leads to the formation of clarifier and other sludges, which must then be disposed of as waste. Landfill leachate must also be treated, and once again this process can contribute to sludge formation. In this way, a complex residue/waste system results, and the solution of one problem results in the formation of secondary waste [78].

5.3.2. Production-Oriented Management

Residues and wastes can arise in more than one environmental area of production. Concepts for residue management must therefore extend over many such areas. Disposal of residues as wastes, considered alone, even if carried out in an appropriate, environmentally compatible form, has the disadvantage that the waste problem is dealt with at the end of the production chain; there may also be ecological and economic drawbacks associated with this arrangement. These end-of-pipe technologies are generally more expensive than upstream measures (see also → Production-Integrated Environmental Protection).

Waste management in production is a planned, purposeful technique including three elements [79]:

1) Analysis of wastes and waste production by type and quantity in the production chain. This makes it possible to device an "avoidance" concept.
2) Environmentally compatible technical measures (disposal) or new solutions for the waste problem on the basis of the analyses.
3) Monitoring of the practices implemented.

Today, the principle of waste avoidance is incorporated in new process development (efforts to avoid generating residues or decrease the quantity generated). Existing processes are also studied and reconfigured if appropriate. This is the principle of production-integrated environmental protection. Another way to implement this principle is to utilize residues; the relative priority of utilization and waste disposal has to be established case by case. The ecological and economic effects of utilization have to be analyzed; one question is whether reclamation of materials may harm the environment more than it helps, for example when recovery processes consume disproportionate amounts of energy. Again, many reclamation processes generate wastes, wastewaters, and waste gases, which in turn must be disposed of. Finally, a reclamation process cannot be repeated indefinitely, and utilization has to be viewed from the market standpoint. Accordingly, it is necessary to clarify in advance whether the reclaimable substance is required and accepted by downstream processors or consumers. Problems can arise when the new material obtained by reclamation is of lower quality than the primary product. If there is no market, even the best reclamation concept is doomed to fail. Even a viable reclamation concept can come to nothing if the level of acceptance changes abruptly. The decision between utilization and disposal must also take account of ecological balances prepared by scientifically sound and objective methods. This statement also holds for thermal utilization, which ranks equally with recycling of substances: A residue is thermally utilizable if its combustion produces

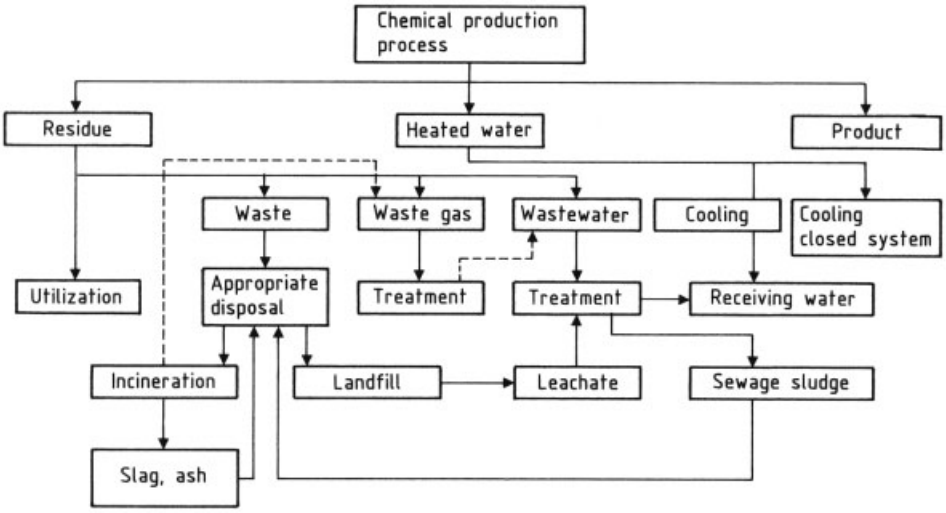

Figure 12. Principles of chemical production processes

more heat than the sum of the losses experienced. The difference between the heating value of a substance and the unavoidable losses in combustion can be recycled to the production process in the form of steam and electric power. This energy replaces primary energy sources such as coal, petroleum, and natural gas. If the energy balance is zero or negative, however, thermal treatment becomes a form of disposal.

Because residue reduction and reclamation cannot lower the quantity of production waste to zero, environmentally sound disposal will remain the basis of waste management for the foreseeable future. Disposal takes the form of incineration or landfilling. Each of these activities today meets high technological and safety standards and ecological concerns are amply addressed. Emissions into air, water, and soil have been cut to a very low level. A modern disposal concept also includes analysis of the present quantity, makeup, and properties of the waste and waste streams along with short-term projections. Whenever possible, secondary wastes such as ash and slag produced by incineration are also utilized, and the heat produced is reclaimed in the context of energy recycling.

5.3.3. Product-Oriented Management

When products of the chemical industry are processed and consumed, they are partly converted to waste. A reduction of such product wastes is necessary and can be achieved through recycling of product wastes or the formation of closed loops [80]. The chain of responsibility from producer via downstream processor and consumer must also be completed under a "holistic marketing" concept [81]. This means considering what happens to the material after use. The technical reclamation measures have to be matched to each type of product waste.

Recycling can involve recovery of raw materials by chemical cleavage or reclamation of materials with no alteration of chemical structure. It is necessary to verify that necessary cleaning, transport, and energy do not lead to ecological impacts elsewhere. Reclaimed product wastes should always fill gaps in the market and help to save costly raw materials; they should not be justified solely by recycling for its own sake. Again, market conditions must not be left out of account, and the technical constraints on the reclamation of product wastes must be taken into consideration. Plastics wastes, for example, are not indefinitely recyclable without change in composition, because the recycling processes alter the molecular structure. The structural changes

in turn affect macroscopic properties of the plastic, such as toughness and strength. The recycling system takes the form of a cascade ("downcycling") that ends in the irreversible transformation of matter to energy by combustion with utilization of the energy, or disposal in landfills.

5.4. Disposal Measures

5.4.1. Logistics

Analytical studies are used to monitor and control production processes in the chemical industry and to characterize chemical products for their later use. The same holds for wastes and waste-disposal processes.

Any residue that is not utilized in the production plant is characterized by an extensive list of chemical and safety data. This information is used in assessing the behavior of the waste during shipping, storage, and disposal, and also serves as a guide for selecting the appropriate disposal method. Studies of the physical and chemical properties of wastes are performed and their exact compositions are determined. Data on the production process that gives rise to the wastes are also integrated into the process. A product-by-product breakdown of waste quantities is obtained on the basis of a production registry. In this way, when production changes are planned, the resulting changes in waste streams can be predicted as well.

When wastes are to be disposed of in landfills, a variety of parameters are ascertained by analysis of the original substance and analysis of the aqueous eluate. Other studies are related to landfill emissions such as leachate.

Some special parameters (e.g., flash point, heating value, water content) must be determined for wastes that are to be disposed of or thermally utilized by combustion. All data for a given waste are documented in a multi-page survey form. The plant manager is responsible for the correctness of the information. After thorough checking, including a check for completeness of analysis data, the necessary pretreatment and the disposal route (including the proper transportation method) are decided on by a cental corporate office. Disposal and transformation are only carried out when official approval has been obtained for each individual case. Computer-based systems are used to document all analytical data on wastes, calculate disposal costs, and provide data to the regulatory agencies.

5.4.2. Waste Combustion

Hazardous Wastes. Incinerators are used for solid, pastelike, and liquid wastes that cannot be placed in landfills because of environmentally relevant constituents, provided they can burn without exploding. Such wastes are essentially organic production residues, vegetable and animal fat products, herbicides and pesticides, petrochemical and coal products, organic solvents, dyes, paints, adhesives, putties, and resins. To avoid harmful emissions when these products are combusted, the flue gases must be cleaned.

Before initial delivery, each waste must be examined as to its relevant physical and chemical data. The purpose of this analysis is to make certain that the emission limits as well as occupational health and safety and fire-protection standards are not violated upon combustion. The results of this investigation make it possible to classify wastes, that is, to group them with respect to the combustion process and thus achieve optimal furnace conditions.

Widely used combustion systems include rotary kilns, fluidized-bed furnaces, and multiple-hearth furnaces. A number of wet and dry processes have been devised for cleaning the flue gases. Hot flue gases are cooled and their energy is recovered and used to raise steam before the cleaning stage.

At BASF, flammable solid, pastelike, and liquid residues are combusted in eight furnaces. Each combustion unit consists of a rotary kiln with afterburner chamber and a steam boiler. The superheated 18-bar steam from units 1 to 6 is fed into the BASF plant network. In units 7 and 8, a higher-value steam is generated with an efficiency of ca. 74 % and supplied to a back-pressure turbine, where it is expanded from 43 bar to 5 bar. To utilize the heat of the flue gases between 300 °C and 180 °C, a waste-heat boiler was installed to raise 5-bar steam. Electric power and 5-bar steam are fed into the respective plant systems.

Three flue gas cleaning units were built for the eight rotary furnaces. Each unit consists of

an electrostatic filter and a three-stage scrubber. By 1996, these units will be retrofitted with catalysts to lower NO_x emissions and degrade dioxins. The plant combusted ca. 150 000 t of waste in 1992; on reconstruction, a combustion capacity of 175 000 t/a will be available after 1996 [82].

The Hoechst waste combustor at the Höchst plant has a capacity of ca. 45 000 t/a. The system also consists of rotary furnace, afterburner, electrostatic filter, and flue-gas scrubber. A coke filter has been retrofitted to cut dioxin emissions, and an SCR unit is attached to the filter to reduce NO_x. Scrub liquor is forwarded to a wastewater treatment system for heavy-metal precipitation and then to the plant's biological treatment facility.

Bayer operates a number of similarly engineered waste combustion units. At the Dormagen plant, a further combustor for solid and liquid wastes has been placed in service. It consists of a rotary furnace, afterburner, waste-heat boiler, and gas scrubber. The plant incorporates a condensation-type electrostatic filter specially developed by Bayer [83] and an SCR unit for selective catalytic reduction of nitrogen oxides (with ammonia) and for degradation of dioxins in the tail gas (see Fig. 13).

Sewage Sludges. At BASF, ca. 100 000 t/a of sludge (dry basis) is dewatered to ca. 40 % dry content. This consists of excess sludge from the biological wastewater treatment system mixed with coal and ash (filter aids). The filter cake is combusted in fluidized-bed units with a positive energy balance. The heat liberated by combustion in the furnaces is used to raise steam. Flue-gas cleaning consists of dust removal with an electrostatic filter and wet scrubbing [82]. The multiple-hearth sludge combustion method used by Bayer is similar (see Fig. 14). Waste gases and vapors from the multiple-hearth furnace are thermally treated in a post-combustion chamber with addition of liquid wastes [84].

Hoechst has put a two-stage sludge combustion system in service, with a capacity of 130 000 t/a. The combustion unit is a fluidized-bed furnace consisting of the fluid-bed space proper and a separate, cylindrical secondary reaction zone operated at 850 °C (see Fig. 15). The oxygen content in the fluidized bed is held between 1 and 3 vol % to lower the NO_x concentration. Both carbon monoxide and organic carbon are burned in the secondary reaction zone; preheated secondary air containing 6 – 7 vol % oxygen is injected at the head of the furnace. In this way, fuels that contain nitrogen-containing compounds (e.g., protein compounds in sewage sludge) can be burned with a much lower NO_x formation rate, eliminating the need for a flue-gas denitrification unit [85]. Flue-gas scrubbing liquor is recycled, with a substream going to a wastewater treatment unit. Heavy metals in the scrub liquor are removed there in a multi-stage process:

1) Two neutralization stages with caustic soda (precipitates sparingly soluble metal hydroxides)
2) Addition of sodium sulfide in neutralization stage (precipitates even more sparingly soluble metal sulfides)
3) Flocculation of heavy-metal precipitates with iron(III) salts and polyelectrolytes; settling and filtration

The product water is fed to the biological wastewater treatment facility.

5.4.3. Landfill Disposal of Wastes

Landfill design, the technical and organizational aspects of landfill operation, and the requirements on wastes for landfill disposal are regulated in detail by TA Abfall (an administrative regulation under the German Waste Disposal Act). The focus is protection of groundwater, which is achieved by the use of a landfill foundation seal. This feature can consist of a natural seal (e.g., a bed of clay) or an artificial liner (e.g., plastic sheeting). Leachate formed above the liner is removed through drains, collected, and forwarded to treatment. An example with an artificial seal is the Hoechst AG sewage sludge landfill (see Fig. 16). Leachate is led away via drain pipes and collected for treatment. Another landfill operated by Hoechst, for production wastes, uses a 40 m underground clay bed as its foundation seal. Leachate and precipitation are collected and sulfates are removed in a pretreatment system in which sulfate is precipitated as gypsum. The treated leachate is then forwarded to biological treatment.

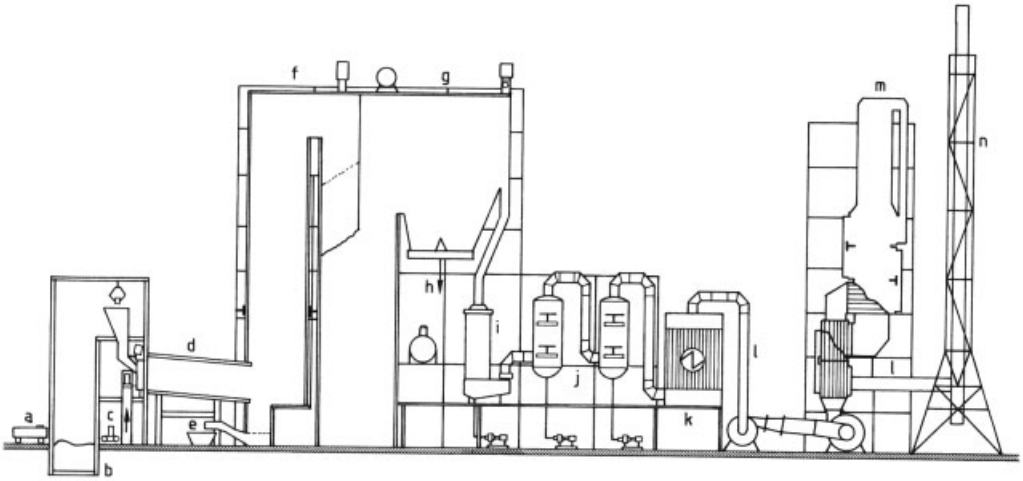

Figure 13. Bayer hazardous waste incinerator
a) Tanks for liquid residues; b) Waste bin for solid waste; c) Drums for solid waste and sludge; d) Rotary kiln; e) Slag discharge; f) Post-combustion chamber; g) Heat recovery (boiler); h) Ash discharge; i) Quench; j) Rotational scrubber; k) Condensation electrostatic precipitator (ESP); l) Exhaust vent; m) Selective catalytic reduction (SCR); n) Stack

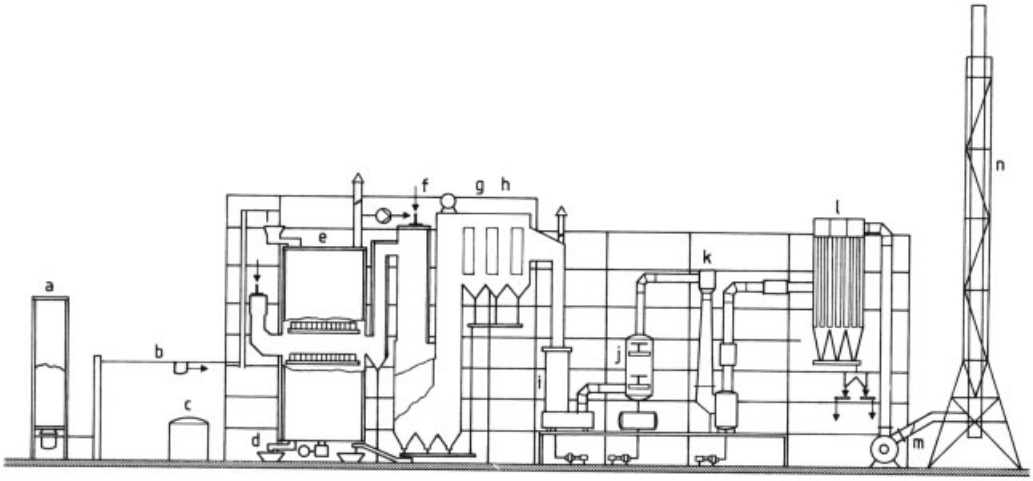

Figure 14. Bayer sewage sludge incinerator
a) Sewage sludge bin; b) Sewage sludge transport system; c) Liquid residues; d) Ash discharge; e) Multiple hearth furnace; f) Post combustion chamber; g) Flue-dust discharge; h) Heat recovery (boiler); i) Quench; j) Rotational scrubber; k) Jet scrubber; l) Adsorber bag house; m) Exhaust vent; n) Stack

For its Leverkusen landfill, Bayer devised new intermediate and cap seals. Textured plastic webs functioning as space-saving intermediate seals form an effective barrier to prevent precipitation penetrating into the wastes until the cap is placed. Each intermediate layer has its own drain system connected to a water interceptor and monitoring system. The construction of the intermediate seals began in 1989. Such seals or equivalent measures, as well as the construction of inorganic barrier layers to protect the slopes and top surface, constitute a dependable multibarrier system. All contact of the waste in the landfill with the environment is blocked. This design ensures ecologically compatible and safe deposition of wastes for the long term [86].

The BASF landfill on the island of Flotzgrün in the Rhine has a double liner of plastic webs, insuring that all leachate is collected and delivered to treatment. The effectiveness of the liner

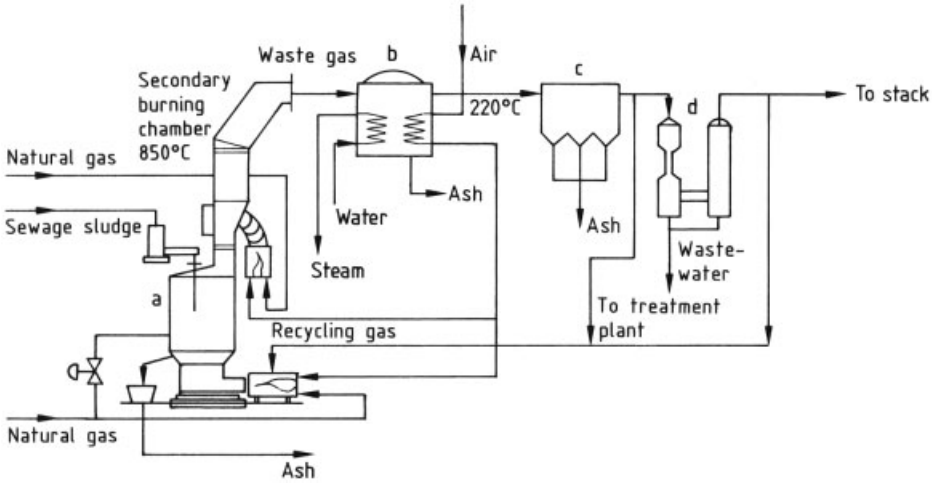

Figure 15. Hoechst sewage sludge incineration plant
a) Fluidized-bed reactor; b) Boiler; c) Electrostatic precipitator; d) Off-gas scrubber

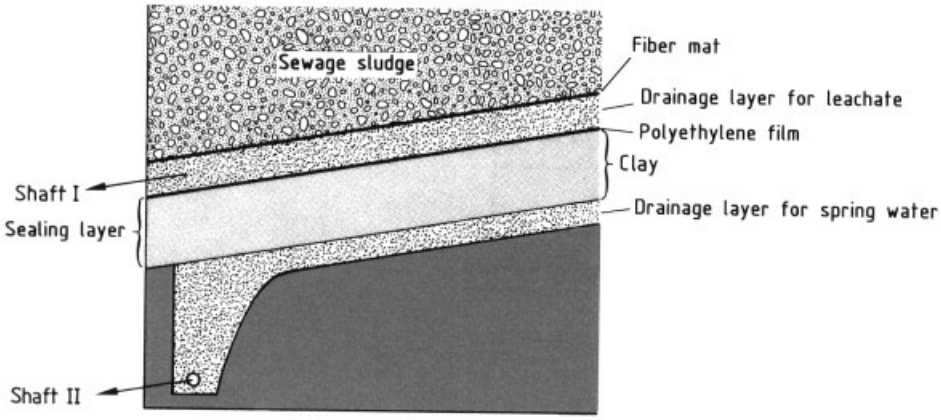

Figure 16. Landfill site with sealing layer for leachate

is continuously monitored, and leaks can be repaired [87].

5.4.4. Asbestos Disposal

Solvay Umweltchemie has devised a process for disposing of asbestos (the Solvas process). Asbestos with a particle size of ≤ 5 mm is reacted with hydrofluoric acid in a batch operation, yielding fluorides and hexafluorosilicates of the elements present in the asbestos-containing material. Crocidolite (blue asbestos) and white asbestos differ in reaction time. After filtration, the solids are neutralized with milk of lime and the filtrate is recycled. Neutralization yields a product consisting of CaF_2, oxides, and metal hydroxides of the elements present in the starting material, and silicate-containing compounds. This end product can be used as an additive in the manufacture of concrete blocks and as a flux [88–90].

5.5. Utilization of Product Wastes

5.5.1. Plastics Recycling

Plastics to be recycled after use must be collected in the purest possible form. The recyclable portion of the waste must be separated into classes of valuable substances or individual substances. The fractions that result can be

recycled for valuable constituents, recycled for feedstocks, or thermally utilized, depending on the type of plastic.

Recycling of Polypropylene. At its Knapsack site, Hoechst has erected a recycling plant for polypropylene with a capacity of 5000 t/a of PP recyclate (Fig. 17). Recycling makes it possible to start with pure PP fractions from used parts (such as those from automobiles) and modify them into "typeware." The material must be collected, ground, and sorted on a decentralized basis. The ground recycled parts are mixed with virgin PP powder, additives, and reinforcements, processed in an extruder, and thus converted to products that meet the required specifications [80].

Recycling of PVC Processing Wastes. A recycling plant for poly(vinyl chloride) wastes has been built at the Hoechst, Gendorf site. The principal feed is rigid PVC film trimmings from film processors. The plant has a capacity of 4000 t/a. The recycling line begins with grinding of the PVC wastes and removal of foreign substances such as paper, wood, and other types of plastic. At the same time, the PVC recyclate is washed and reground. After these steps, the PVC is subjected to final grinding, formulation, and conversion to new raw material for PVC film. In the same plant, this secondary feed is converted to PVC film for nonfood packaging and technical applications. The purpose of this prototype plant is to gather experience for future plants at other locations [80].

Recycling of Polyoxymethylene. The recycling techniques used up to now (regranulation and compounding) can damage polyoxymethylene. The consequences are separate specifications and application restrictions for recyclates. With a new process from Hoechst, these drawbacks can be avoided. The material recovered here is not the polymer itself but the monomer, chiefly trioxane, which is obtained by chemical degradation of POM wastes. The result of this recycling process is always new product. Only about 2 % of the POM recycling feed (consisting of stabilizers) becomes waste; this is combusted. The prerequisite for the process is that used plastic parts and processing residues be collected in pure form. A pilot plant is now in operation and a production unit is in the planning stage.

Recycling of PET Wastes. Hoechst Celanese has built a 70 000 t/a plant at Wilmington, Delaware, to recycle poly(ethylene terephthalate) wastes by methanolysis.

The glycolysis process is used in a recycling system (capacity 5000 t/a) for PET wastes at the Offenbach plant of Hoechst.

Recycling of Used Plastics. BASF is constructing a 15 000 t/a pilot plant for recycling of mixed, unwashed used plastics into feedstocks (Fig. 18).

The conversion of mixed plastics to petrochemical feedstocks takes place in a multistage melting and degradation process. In the first conversion stage, the used plastics are melted, and dehalogenated to protect downstream equipment from corrosion. Hydrogen chloride evolved in this stage is absorbed and fed to the hydrochloric acid plant. Gaseous organic products are compressed and can be utilized as feeds to a cracker.

In the next stages, the liquefied plastics are further degraded and fed to distillation towers. The distillates include naphtha, olefins, and aromatics, which are used as feedstocks in BASF's petrochemical plant. The high-boiling bottoms are converted to synthesis gas [91].

5.5.2. Refrigerant Recycling

In 1989, Hoechst decided that its output of perhalogenated chlorofluorocarbons (CFCs) should be phased out for environmental reasons and stopped by the end of 1994. The ozone-damaging CFCs are being replaced with R134a, whose contribution to the greenhouse effect is 90 % less than that of CFCs. The new product, like the earlier ones, is integrated into a two-stage recycling system.

Recycling of used CFC refrigerants was introduced in 1986 – 87. After collection by dealers, the essential processing steps are mechanical purification and distillation.

Refrigerant mixtures that cannot be separated by distillation, together with CFC still being delivered to Hoechst after production shutdown, are subjected to secondary recycling by a process that is patented worldwide. The refrigerants

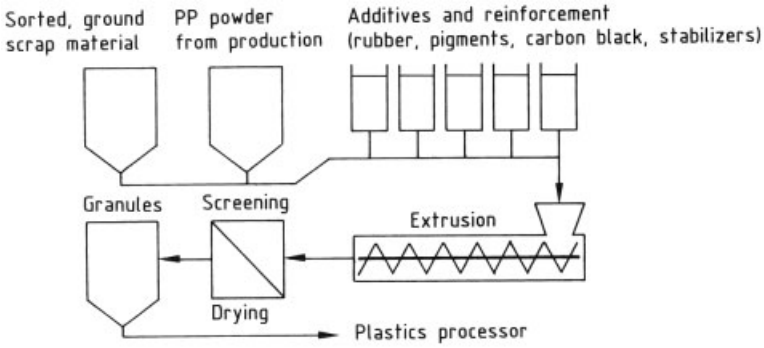

Figure 17. Recycling of polypropylene at Hoechst

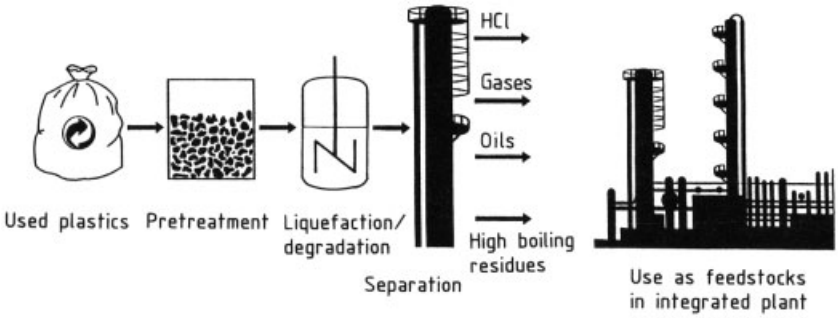

Figure 18. Recycling of used plastics into new feedstocks (BASF)

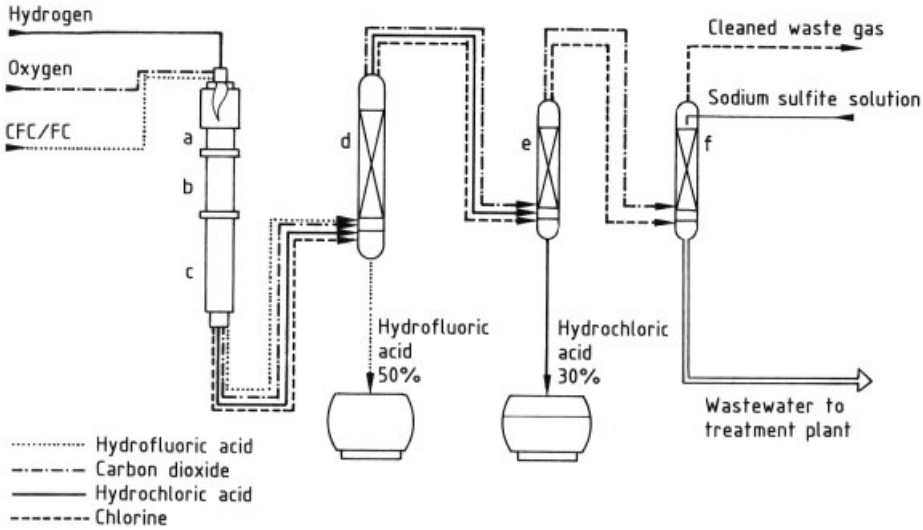

Figure 19. CFC/FC cracking plant with HCl and HF recovery (Hoechst system)
a) Combustion chamber; b) Cracker; c) Absorber; d) HF recovery; e) HCl recovery; f) Scrubber

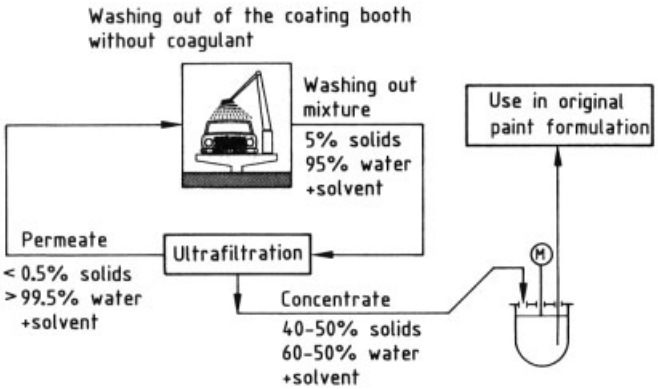

Figure 20. Recycling of paint overspray in automotive production line coating

are cracked in a thermal reactor, and hydrochloric and hydrofluoric acids are recovered from the reactor outlet gas (Fig. 19). These acids are used as feeds for other production processes [92].

5.5.3. Recycling of Used Packaging Materials

In a pilot project of the Verband der Chemi-schen Industrie, a facility for disposing of chemical drums was erected in the Hoechst Ruhrchemie plant. After pretreatment, drums with capacities of $\geq$120 L are cleaned, dried, and examined for damage and cleanliness. Clean drums are forwarded to recycling. Damaged drums larger than 120 (60) L as well as small containers (<120 or 60 L) are ground, washed, and dried. The materials, in the form of pure steel scrap or sorted plastic chips, are then available for recycling [93].

In 1991, BASF placed a central drum-cleaning plant in service. The next year, some 330 000 empty drums were cleaned in three automatic cleaning lines. The drums varied in size, material (plastic and steel), and design (open-head and tight-head). Two lines are run with water (with additives) and one with solvent [82].

5.5.4. Paint Recycling

Product development in water-based paints must consider how paint residues are to be recycled in the user's painting facilities. In certain paint systems, such as auto body paint booths, overspray is caught in circulating aqueous liquors and reconcentrated by ultrafiltration (see Fig. 20). Concentrate solutions are added to the appropriate paint formulations. In a special stove-drying paint system, the material recycling rate is ca. 95 %. The special membranes needed for the ultrafiltration step were developed by Hoechst subsidiaries Herberts and Hoechst Celanese Corporation, together with plant construction specialists.

5.6. Results of Waste Management

The results of waste management are characterized by two areas. The primary one relates to avoidance, reduction, and recycling of residues generated in production. This is achieved by means of production-integrated environmental protection (→ Production-Integrated Environmental Protection), leading chiefly to avoidance or reduction of waste.

The second area, waste disposal, exists because no chemical plant can operate without generating waste. Techniques used here include landfill disposal when appropriate, as well as waste combustion with energy recovery and modern flue-gas cleaning. Disposal is described in terms of the type, quantity, and disposal mode of the wastes. The following information on these processes has been provided by BASF, Bayer, and Hoechst:

BASF. The quantity of wastes combusted with energy recovery in 1983 was 94 000 t. This quantity rose steadily in later years, reaching 149 000 t in 1992. A decline to 143 000 t was recorded in 1993 (see Fig. 21).

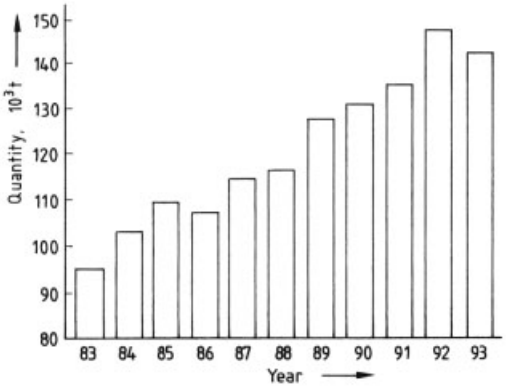

Figure 21. Waste incineration at BASF, 1983 – 1993

Figure 22 shows the quantities of waste disposed of in the Flotzgrün landfill from 1983 to 1993. These wastes are classified as process wastes, wastewater treatment sludges, and building rubble and construction materials. The decrease in quantity since 1989 is attributed to recycling of building rubble and avoidance and recycling of process residues. Since 1992, all sludge from wastewater treatment has been combusted, so in future this source will contribute only ashes to the landfill. Other wastes deposited in the landfill in 1993 were 67 000 t of process wastes, compacted trash, trash from plant cleaning, and wastes from sludge combustion, together with 29 000 t of building rubble and construction materials [87].

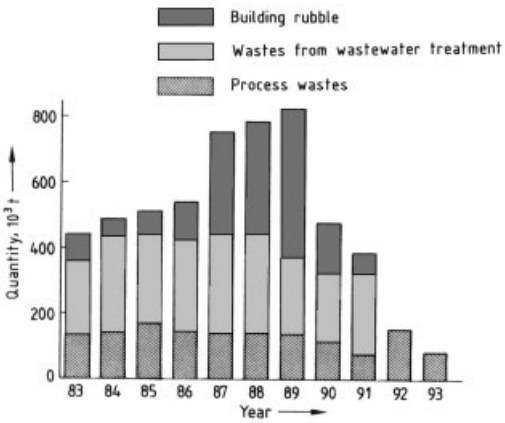

Figure 22. Quantities of wastes deposited in the Flotzgrün landfill (BASF)

Bayer. The total amount of wastes generated in 1993 was 667 000 t. These quantities represent a reduction from the years 1981 – 1992. Figure 23 gives a breakdown of wastes with total quantities for the period from 1981 to 1993. The 7 % increase in waste produced between 1981 and 1988 was essentially due to a gain in production (nearly 16 % in the same period). Various practices introduced in production plants since 1988 have led to a reduction from 506 000 t to 339 000 t of production wastes in 1993 [94].

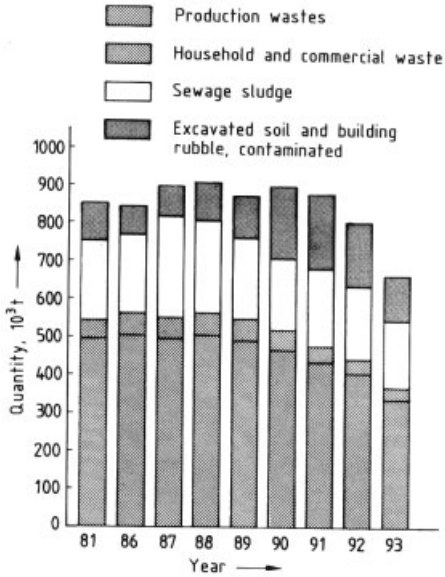

Figure 23. Waste quantities (Bayer)

Hoechst. The year 1993 saw a total waste production of 348 000 t, comprising 42 000 t of household-type waste, 173 000 t of sewage sludge, and 133 000 t of special and hazardous wastes. The last figure includes 44 000 t of contaminated building rubble and excavated soil. Figure 24 gives a breakdown by these same three waste categories for the period 1983 – 1993. The marked decline in total wastes generated and also in hazardous wastes can be attributed both to the shutdown of certain production operations and to production-integrated environmental protection [95].

The total waste quantity for 1993 (348 000 t) breaks down as follows:

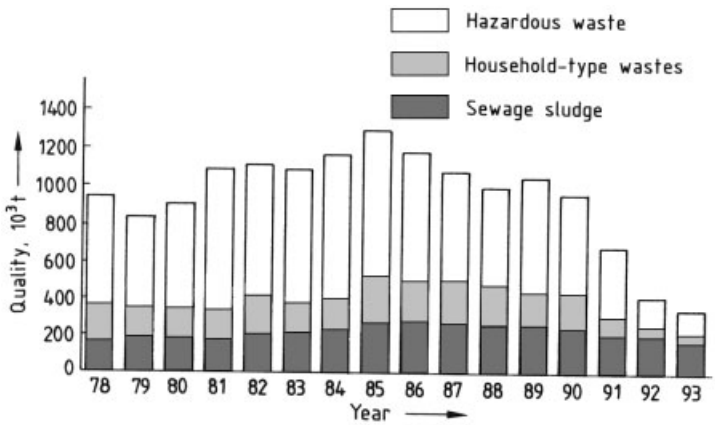

Figure 24. Waste quantities (Hoechst)

Sewage sludge	55 %
Building rubble	16 %
Distillation residues	9 %
Solvents, solvent mixtures, mother liquors	7 %
Residential waste	5 %
Slag, ash	2.5 %
Plastics and polymer wastes	1.4 %
Diverse wastes	4.1 %

Waste disposal in 1993 breaks down as 89 000 t combusted and 259 000 t deposited in landfills.

6. Prevention of Waste

6.1. Introduction

Prevention of waste production can be achieved by two means: dispensing with production and consumption or technological waste prevention measures (see Fig. 25). Dispensing with production or consumption is the most dependable way to prevent environmental burdens due to human activity. Products that are not consumed need not be produced and thus cannot have an impact on the environment [96].

In contrast, the goal of technological waste prevention is to satisfy needs with minimal environmental impacts. Impacts result from consumption of resources and generation of wastes and emissions. Other prevention measures include cutting emissions to the atmosphere and water bodies, reducing the potential harm from emissions, increasing the service lives of products, and improving their recyclability. This task is of particular importance where all possibilities for dispensing with production and consumption have been exhausted, because products are required to at least meet basic human needs. Waste prevention is influenced by legislation but also by social and economic factors.

Technological waste prevention is broken down into primary and secondary waste prevention. Primary measures aim to reduce both the material stream and the material flow rate [97] by means of planning, development, and optimization of products. Here "products" refers to products for consumption and use, means of production, and infrastructures (buildings, transportation, and communications systems). The purpose of secondary waste prevention is to avoid disposal of products, production wastes, and waste materials by waste treatment or landfill deposition. Secondary waste prevention thus means recycling, the goal being to return the products to the use cycle at the lowest possible cost in material and energy terms. There are several forms of recycling: modifying or converting product while it is in use (re-use), recovering and reprocessing used products or waste materials after their useful life, and recycling production wastes and media back to the production process, with or without passing through a preparation process [98].

Integrated technical environmental protection cannot be analyzed in the same way as technical waste prevention (see Fig. 26) but does pursue the same goals. Integrated technical environmental protection is broken down according

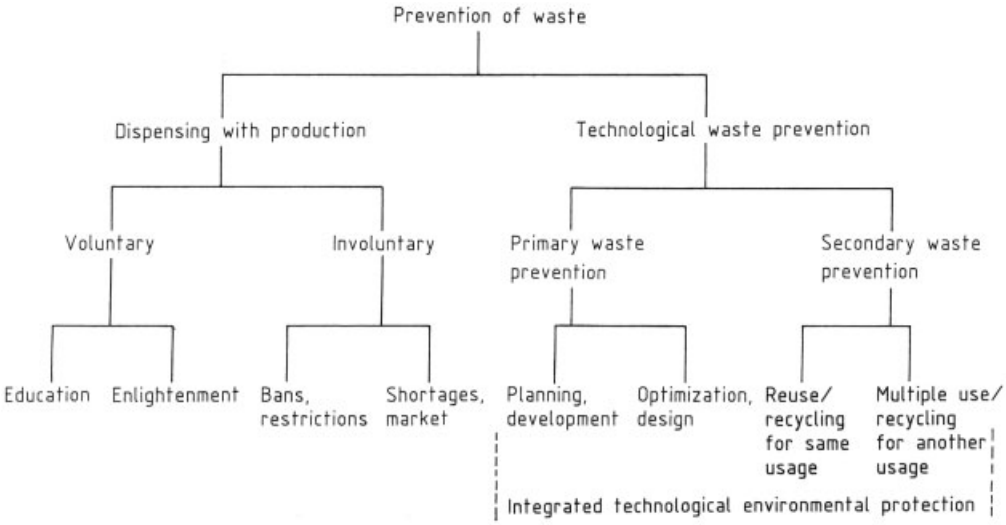

Figure 25. Breakdown of waste prevention [96]

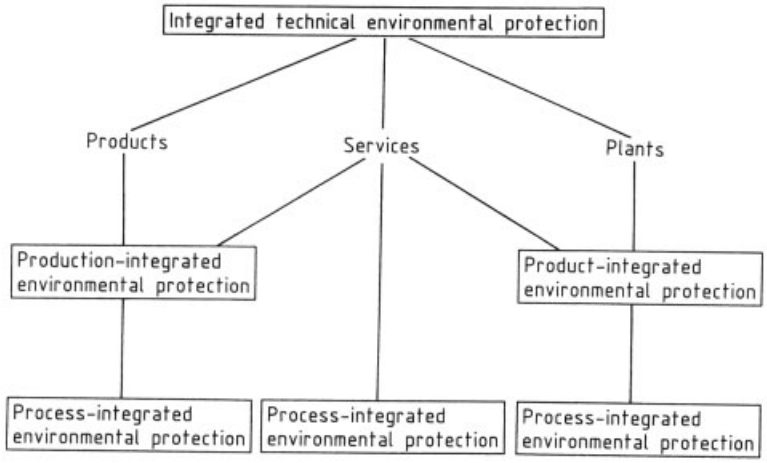

Figure 26. Structure and applications of integrated technical environmental protection [99]

to the system that is at the focus; i.e., product-integrated, production-integrated, and process-integrated environmental protection [99].

Product-integrated environmental protection [100] seeks to minimize use-specific environmental impacts over the entire career of the product (cradle to grave i.e. product line) [101]. This begins with extraction of resources from the earth's crust or biosphere and extends to the reintroduction of wastes and emissions into the natural environment. This approach is the most far-reaching under the different categories of integrated technical environmental protection and is the basis for unambiguous assessment of net reductions in environmental burdens.

The goal of production-integrated environmental protection [102] is to minimize the sum of environmental impacts within a company or a plant complex considered as a subregion of the company. Accordingly, the system boundaries are the boundaries of the company or plant complex. This approach has been developed by the chemical industry, in particular, as an instrument for economic and ecological optimization even where material streams are strongly cross-linked [103]. It can also handle multiple product lines in a company producing a range of prod-

ucts [101]. The procedure is also specified in European legislation (the "ecological audit" system) but cannot guarantee in every instance that a global optimum has been found for the products under consideration.

Process-integrated environmental protection [104] is a component of both product-integrated and production-integrated environmental protection and thus provides a link between these approaches. The objective is to minimize the sum of environmental burdens due to a process. This local optimum can contribute to the global "cradle-to-grave" optimum; that is, the process parameters that yield the global optimum may differ from those corresponding to the local optimum.

The technological basis for these integrated material- and process-oriented approaches comprises knowledge of materials, process and manufacturing engineering, and design engineering. Along with these technical aspects, instruments are needed to objectively assess the environmental burdens caused by the objects of study or anticipated for the future. The techniques of life cycle assessment, still being developed and standardized, will play a role here (see Chap. 4) [105].

Integrated technical environmental protection, like technical waste prevention measures, covers all technical actions and methods used to minimize environmental burdens due to products, services, plants, and processes. This is a holistic, systems-analysis approach that takes account of all interactions resulting from the individual technical measures. The overall aim is to ensure that products and services meet demands and carry out their functions, while raw material and energy consumption as well as environmental burdens are minimized.

6.2. Distinction between Primary and Secondary Waste Prevention

The distinction between primary and secondary waste prevention is equivalent to the distinction between legislatively mandated prevention and recycling [106]. An unambiguous distinction can only be derived by beginning the analysis with the process. As Figure 27 shows in simplified form, prevention is characterized by the fact that the masses m_0 inlet to the process are reduced by an amount m_p. In Figures 27 and 28, what is shown as a mass may actually be a sum of masses. If decreasing the original input m_0 entails the use of additional resources m_A, the actual prevention m_p is the difference between the sum of all actual inputs and the sum of all original inputs m_0. In this argument and those that follow, it is assumed that the mass m_Pr and the quality of the process product are the same in each case.

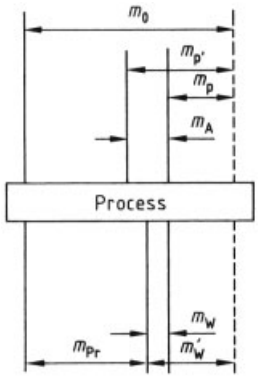

Figure 27. Simplified diagram of process-oriented waste prevention m_p

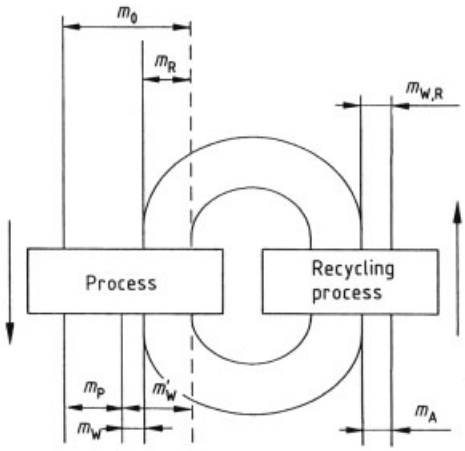

Figure 28. Simplified diagram of process-oriented waste abatement m_R (with recycling)

The recycling shown in Figure 28 leads to the replacement of some of the original input m_0 by secondary materials m_R generated in the recycling process. The mass input to the process is the same as the original input m_0 in Figure 27. When the system boundaries are expanded, no distinction is made between the re-

duction in mass m_p due to prevention and recycling; that is, if the entire production system is examined from an external viewpoint, primary and secondary waste prevention practices can cut the original amount of waste m'_W to the same value m_W. Apart from the closed- circuit aspect recycling involves at least one recycling process, which may be merely a transportation process. The input of matter and energy to this recycling process is not considered in the analysis of the process itself. In contrast to primary waste prevention, here it is not possible to state whether the net effect of recycling is a reduction in resource use and thus in environmental burdens [107].

Whether primary or secondary waste prevention is under consideration, process-oriented prevention must be distinguished from production- and product-oriented prevention. The avoidance of wastes and emissions from a given process can, but does not necessarily, lead to a reduction in the amount of wastes and emissions generated in the plant or throughout the product life cycle. The process analysis is the only certain way to distinguish primary from secondary waste prevention, or, in general, prevention from recycling.

6.3. Waste Prevention and Reduction of Environmental Burdens

The requirement to meet demand or carry out functions results in "profiles" of required product service properties. While a reduction in demand can be achieved by renunciation, technical waste prevention is based on meeting constant demand and thus unvarying requirement profiles for product properties. In other words, all further discussion relates to comparing products having the same benefit [108].

The environmental burdens produced by a product, directly or indirectly, throughout its life cycle from time t_0 when raw materials are extracted from the Earth's crust or biosphere up to t_∞ when wastes and emissions are reincorporated in the environment, are symbolized by $\mathrm{EB_{Pr}}$ in Equation (6.1). For each substance i, the ecologically weighted primary resource consumption M_{PR} (material and energy) and the ecologically weighted mass $\mathrm{M_{WE}}$ of wastes and emissions generated are summed over the life cycle. Ecological weighting takes into account the potential impacts due to the consumption of various resources and the generation of various wastes and emissions. As Equation (6.2) shows, the ecologically weighted mass M is the product of a weighting factor b and the mass m. The weighting factor b is formed similarly to the categories of the impact assessment of life cycle and takes account of, for example, the difference between a nearly environmentally neutral mining spoil and a far more relevant galvanic mud containing cyanide. Energy is treated similarly to mass, or the energy is converted to a corresponding mass of the primary energy resource [109].

Because any consumption of material and energy is equivalent to environmental burdens, there is no product that is absolutely environmentally friendly. For a given requirement profile the sum of the impacts on the environment (including those on humans) must be minimized by prevention measures. When a component or product can be made with more than one construction material or compound that meets the requirement profile, the material or compound that causes the least environmental burdens over the product life cycle should be chosen [99]. In contrast to Equation (6.1), where burdens are related to the product alone, the identification of the ecologically favored variant must allow for the service life $(t_e - t_n)$ or the frequency with which functions are carried out (Eq. 6.3). The sum of life cycle environmental burdens divided by the service life describes the material flow rate to be reduced [109]. The environmental burdens due to a product is not an absolute quantity, because the actual impacts cannot be quantified. For this reason, only the difference in use-specific environmental burdens (UEB) between two or more products can be determined (Eq. 6.4) in order to ascertain which product is ecologically more favorable and thus more environmentally benign. The attribute "environmentally friendly" is relative, not absolute.

$$\mathrm{EB_{Pr}} = \sum_{t_0}^{t_\infty} \sum_i (M_{\mathrm{PR}} + M_{\mathrm{WE}})_{\mathrm{t,i}} \tag{6.1}$$

$$M_i = b_i m_i \tag{6.2}$$

$$\mathrm{UEB_{Pr}} = \frac{\mathrm{EB_{Pr}}}{t_e - t_n} \tag{6.3}$$

$$\Delta \mathrm{UEB} = \left(\frac{\mathrm{EB_{Pr}}}{t_\mathrm{e} - t_\mathrm{n}}\right)_1 - \left(\frac{\mathrm{EB_{Pr}}}{t_\mathrm{e} - t_\mathrm{n}}\right)_2 \tag{6.4}$$

In Equations (6.1) – (6.4), EB denotes environmental burdens; M is the ecologically weighted mass; UEB is use-specific environmental burdens; b is the ecological weighting factor; m is mass; t_x is time (see Fig. 29); Pr refers to product, PR to primary resources, and WE to wastes and emissions.

Figure 29 illustrates how the environmental character of a product is determined as the ecologically weighted mass M for the product as a function of time over the life cycle [107]. For the sake of simplicity, the ecologically weighted masses of wastes and emissions (M_WE) are set equal to those of the secondary and primary resources ($M_\mathrm{SR}, M_\mathrm{PR}$). In reality they are generally different. In the time from t_0 to t_n, the finished product is manufactured in n process steps. From the primary resources PR and secondary resources SR input to each process step, the process product (shaded) and the wastes and emissions (white) are generated. During the product life, discrete quantities consumed are replaced by continuous consumption of media from primary resources. The wastes and emissions produced during the product life are not shown until the end of the service life t_e in this simplified plot. In the period from t_e to t_m, m process steps are carried out to produce secondary raw materials (SR). Primary resources are consumed in recycling and in subsequent waste processing and disposal. Waste processing includes air and water pollution control measures.

As Figure 29 shows, the life cycle of a product can be divided into four periods: manufacturing (M) ($t_n - t_0$), use (U) ($t_\mathrm{e} - t_n$), recycling (R) ($t_m - t_\mathrm{e}$), and waste treatment and disposal (D) ($t_\infty - t_\mathrm{m}$). Each segment of the life cycle contributes to the total environmental burdens $\mathrm{EB_{Pr}}$ (Eq. 6.5). For the product planner, the terms in Equation (6.6) are so-called ecological mortgages [99], which must be known in order to minimize the total environmental burdens.

$$\mathrm{EB_{Pr}} = \mathrm{EB_M} + \mathrm{EB_U} + \mathrm{EB_R} + \mathrm{EB_D} \tag{6.5}$$

$$\mathrm{EB_M} = \mathrm{EB_{M,PR}} + \mathrm{EB_{M,SR_1}} \tag{6.6}$$

The cumulative ecological burden for production is equal to the sum of environmental burdens due to primary and secondary resource consumption $\mathrm{SR_1}$ (Eq. 6.6). The production of secondary resources is analyzed as if secondary resources were generated not from recycled wastes but from primary resources. In other words, each environmental burden is determined from the ecologically weighted masses of primary resources balanced in i or j processes along with wastes and emissions (Eqs. 6.7 and 6.8). The cumulative ecological burden of recycling is the difference between the environmental burdens of the primary resources consumed and of the secondary resources $\mathrm{SR_2}$ formed (Eq. 6.9). For determining the resulting decreases in environmental burden, the secondary resources obtained by recycling are treated as if they were made from primary resources, not from wastes for recycling. In this way, the cumulative ecological burdens of recycling are the burdens due to recycling minus the burdens that would occur if the secondary resources were made from primary resources (Eq. 6.9) [110]. The cumulative ecological burdens of utilization ($\mathrm{EB_U}$) and waste disposal ($\mathrm{EB_D}$) are calculated analogously to Equation (6.7).

$$\mathrm{EB_{H,Pr}} = \sum_{t_0}^{t_n} \sum_i (M_\mathrm{PR} + M_\mathrm{WE})_{t,i} \tag{6.7}$$

$$\mathrm{EB_{M,SR}} = \sum_t \sum_j (M_\mathrm{PR} + M_\mathrm{WE})_{t,j} \tag{6.8}$$

$$\mathrm{EB_R} = \mathrm{EB_{R,PR}} - \mathrm{EB_{R,SR_2}} \tag{6.9}$$

Table 11 gives a breakdown of practices that can reduce environmental burdens, broken down by segment of the product life cycle and action areas.

The consumption of resources, the generation of wastes and emissions, and the pollution potential can be influenced throughout the life cycle. Measures taken to reduce use of resources are largely the same as those to decrease the amount of wastes produced. The most important life cycle segment for technical waste prevention, however, is planning and development as a component of production and manufacturing, since it is during this phase that the sum of the anticipated environmental burdens is determined.

The service life of a product is affected by planning and design, production, and use of the product. Recycling is a limiting case in that the service life is extended by the re-use of a used product but not by the recycling of a waste product. Recycling of a waste makes it possible to recover materials or other ingredients, and if these

Table 11. Possible ways to reduce environmental burdens over the life cycle of a product [96]

Segment of product life cycle	Ways to cut resource consumption	Ways to cut down wastes and emissions	Ways to reduce potential harm	Ways to increase service life	Ways to increase recyclability
Manufacturing/production Product design Extraction of raw materials Production of material Production of finished product Distribution	High direct yield High efficiency Process control Continuous processing Short process chains Materials selection Lightweight design Miniaturization Logistics Optimal processes	(as under "resources") + High concentrations Low dissipation	Low-polluting processes and feedstocks Renewable resources Process control Additive environmental protection	Performance of function Process safety Appropriate materials Durable joints/fastenings Durable design (stable, wear- and corrosion-resistant) Composite materials Material compatibility Suitable processes	Design for disassembly and recycling Few materials Compatibility of reclaimed waste Few composites Detachable joints/fastenings Identifiability Modular systems
Product use Use Consumption	Low energy consumption Low consumption of water, other utilities	Selection Minimum consumption Proper disposal	Selection Minimum consumption	Appropriate use and consumption Maintenance and care	
Recycling Collection Re-use Rehabilitation Reclamation of components Reclamation of materials Distribution	High direct yield High efficiency Process control Continuous processing Short process chains Logistics Suitable recycling processes	(See under "resources")	Low-polluting processes and feedstocks Process control Additive environmental protection	Gentle handling Gentle cleaning and disassembly Suitable processes	Suitable recycling processes
Waste processing Collection Physical, chemical, thermal processes	High direct yield High efficiency Process control	(See under "resources") Logistics	Feedstocks Process control Additive environmental protection		
Landfill disposal	Volume reduction	(See under "resources")	Immobilization Encapsulation/sealing		

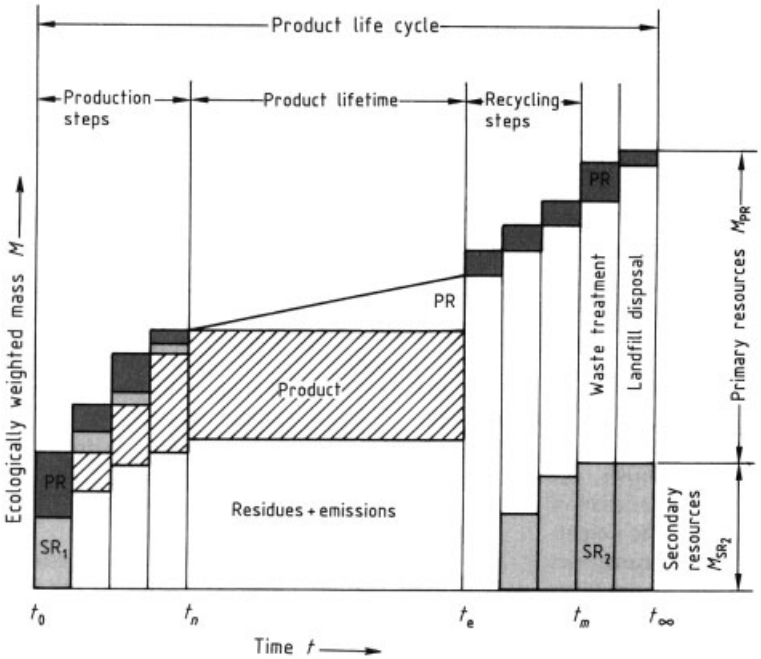

Figure 29. Schematic depiction of ecologically weighted masses M over the life cycle of a product [107]
PR denotes primary resources, and SR secondary resources

are of suitable quality they can lower specific consumption of raw materials for making the product. The recyclability of a product generally can be predicted only in manufacturing, that is, during the design and development phases. An exception is the development of new recycling processes and the optimization of existing ones, which has a beneficial effect on recyclability [111].

6.4. Specific Actions to Reduce Environmental Burdens

The actions listed in Table 11 are specific steps, but by themselves they do not necessarily result in less environmental burdens over the product life cycle. Many of these measures interact with others and may even work against others. Global optimization in the sense of integrated technical environmental protection is necessary to determine how much each specific action is to be stressed.

6.4.1. Decreasing Resource Consumption

In Production. With regard to the consumption of raw materials per unit of end product, it is essential that the feedstocks and intermediates inlet to a process be transformed as completely as possible to the end product and that large fractions of these streams not be lost as residues or byproducts. The ratio of substances inlet to the process to substances present in the end product is called the yield. It is a measure of the utilization of substances entering the process. The higher the yield, the lower the consumption of raw materials and energy per unit of product discharged.

The energy efficiency is the ratio of energy utilized to the quantity of energy contained in primary energy resources. By increasing the energy efficiency, that is, lowering the energy consumption per unit of product, it is possible to decrease both the product burden and the use-specific environmental burdens (such as CO_2 emission). In this context, great importance attaches to improving the energy efficiency in elec-

tric power generation. In the steam or gas turbine used for power production, lengthening the blade cooling is one way to raise the inlet gas temperatures and thus increase the efficiency [112].

Processes must be controlled in such a way that feeds to the process are selectively transformed to the desired products with high efficiency and high yield, while inevitable wastes are suitable for recycling.

This presupposes a detailed knowledge of the process, the existence of a process model, and the availability of suitable instrumentation and control hardware including integrated process control systems. An example is the blast furnace, in which coke consumption has been lowered from 800 kg per tonne of pig iron to under 500 kg/t in the past 20 years as a result of these measure [113].

Batch processes always have an initial phase in which the process parameters are brought to their proper values, and a terminal phase in which the end product is discharged. Additional energy is required for both phases. If the material stream is large enough, batch operation can be changed over to continuous processing, in which the absence of the initial and final phases results in higher efficiencies and better yields.

Lightweight design not only cuts down the amount of resources consumed in making the product, but also reduces environmental burdens if the product has to be transported over long distances or where the product is itself the means of transportation. Lightweight design has a beneficial effect on energy consumption during use of the product.

The task of logistics is to ensure that goods reach the producer, or are delivered from the vendor to the customer, at the proper time and in the needed quantity. Logistics can contribute to waste abatement by route optimization, batch-size optimization, and full utilization of transportation capacity and selection of carrier [114].

Packagings consist of materials in the form of partial or complete coverings needed for the transportation, storage, or use of products as well as to carry product information. Thus superfluous packaging not required for a protective, informational, or transportation function should be dispensed with. Further packaging can be prevented by optimization of the filling quantity and optimization of packaging design, and in some cases by using returnable or multiple-use packagings.

Every process step involves the consumption of energy and materials, and thus the production of wastes and emissions. Whenever the number of operations required to make a given product can be reduced through process development and planning, this shortening of the process chain will result in smaller amounts of wastes and emissions [115].

In Use of the Product. One way product designers can reduce environmental burdens of consumer products is to prepare instructions that are easy to understand and easy to follow.

The most important practice in this area is the selection of products made by environmentally benign methods and featuring low consumption of energy, water, and/or other media. Objective information about the cumulative ecological burdens (mortgages) is a prerequisite for this practice (see Section 6.3).

Careful energy use means, for example, that only the amount of warm water actually needed is used or that heated rooms are only ventilated in a controlled manner.

Another significant potential for waste prevention is following the directions for the use of cleansers, detergents, care products, and other materials.

Good stocking practice can ensure that (in particular) foods, delicacies, and medications are not held past their sell-by dates and thus do not become wastes.

Proper maintenance is important for consumption of resources because it is the key to extending the life of plants, such as chemical industry facilities and machine tools, and of infrastructure, such as buildings and roads.

In Recycling. The opportunities for waste prevention in recycling are the same as in production or manufacturing. The frequently complex nature and rapidly changing composition of waste products, which are often mixed and/or contaminated, is a complicating factor. For this reason, process control in recycling is more difficult than in production processes where primary raw materials of well-defined composition are used. It is also more difficult to model recycling processes and additional instrumentation

and control measures are necessary to give high yields and quality.

In Waste Treatment. Achieving a high direct yield in waste treatment means minimizing the amount of energy and additives that must be put into the waste to make it suitable for landfilling. High energy efficiency is crucial, especially for thermal waste treatment.

As in recycling, the feeds to processing are mixtures with a broad spectrum of constituents, so that process control with low energy and material consumption is possible only with process modeling and appropriate instrumentation and control.

In Disposal. Like waste treatment, deposition in landfills is not a waste prevention measure. If waste volume is reduced, however, the capacity of existing landfills can be better utilized, so that landfill construction and operation entails less consumption of resources per unit of waste.

6.4.2. Reduction of Waste and Emission Quantities

The measures taken to reduce the production of wastes and emissions are largely the same as those used to cut resource consumption. The connection is formulated in the second "theorem" of waste sources [101]: Every resource extracted from the earth's crust or biosphere becomes waste and emissions.

One way to reduce the quantity of waste and emissions in manufacturing and production is to make sure that solid wastes, wastewater, and waste gases have high concentrations of pollutants. High concentrations of emitted substances reduce the absolute quantity of emission media and make it possible to recover valuable substances at low cost or to convert harmful substances to innocuous forms.

Similarly, low dissipation of valuable and harmful substances in the various media lowers the cost of recovery or separation.

After the product's useful life, proper disposal helps diminish wastes and emissions. If used products are sorted by type, recycling or waste processing will cost less than if they are left mixed.

6.4.3. Reduction of Pollution Potential

The use of low-pollutant feedstocks, auxiliaries, and media not only reduces the amount of new pollutants formed in the process but also keeps the quantity of emittable pollutants low. Low-pollution processes are those that have a high direct yield and high efficiency and do not contribute to the formation of new pollutants. The prerequisites for not producing new pollutants are a knowledge of the pollutant-forming mechanism (e.g., de novo synthesis of dioxins [116]), effective instrumentation and control, and design practices that, for example, shorten residence times or promote backmixing.

The use of renewable and degradable feedstocks is a way of cutting or eliminating waste. The use of renewable resources, if they are extracted in a "natural" and sustainable way, can both lower resource consumption and diminish the pollution potential of wastes and emissions.

The polluting potential of emissions can be reduced by separation and treatment of emissions from waste gases and wastewater that is, by measures of the additive environmental protection.

In use, the pollution potential can be reduced by selecting products made by low-pollution processes. Precondition is a comprehensive information.

Except for the use of renewable resources, the measures for lowering the pollution potential in recycling and waste processing are the same as those used in production and manufacturing.

The pollution potential in landfill disposal can be reduced if the substances are immobilized so that the landfill itself does not become a source of hazardous waste. Immobilization also diminishes the risk of pollutants escaping into the environment. This includes encapsulation and sealing of landfills not provided with adequate barriers.

6.4.4. Extending Product Service Life

To have service life, a product must perform the required functions and must afford the specified level of safety in use [117]. A product that does

not meet these requirements could fail prematurely, possibly causing an accident with environmental burdens.

Important manufacturing factors in the service life of a product are selection of materials, type of joints and fastenings, construction and geometry of the product, and nature and fabrication of the product itself [118]. This is a set of interdependent factors.

One way to lengthen service life is to use high-quality materials and composites with resistance to weathering, corrosion, wear, and biological attack. Generally, such materials have a higher cumulative ecological burden for manufacturing than do materials yielding a shorter product life. The selection of a high-quality material reduces environmental burdens only if the gain in service life is greater than the increased burdens due to manufacturing and recycling of the material (see Eq. 6.3).

The same holds for surface coatings (e.g., electroplating, enamel, paint) used to extend the life of a product. The added expense of the surface coating as well as the increased cost of recycling these composite materials must be considered.

The manufacturing technique can strongly affect the qualities of a material or structural element. Practices such as strain-hardening of metals can boost service life without increasing the amount of material used.

Appropriate use or consumption can increase service life during the product service period. If products are employed for purposes they were not intended for or under conditions not provided for in the directions, their service lives can be markedly curtailed.

Care and maintenance are important factors in extending service life during the product use phase.

At the recycling stage, lengthening the service life applies only to products intended for re-use for the same or another purpose. Appropriate practices include careful handling and, if applicable, mild cleaning and disassembly. For products that are not intended for multiple use but are still to be put to new use without any change of form, suitable processing and reconditioning processes must be devised.

6.4.5. Improvement of Recyclability

Measures to improve recyclability may conflict with the extension of service life (e.g., use of composite materials).

Because recyclability has to do with primary waste prevention, the only time when it can be controlled to reduce environmental burdens is at the design and development phases of manufacturing or production. Design for ease of disassembly and processing to improve recyclability [111], [119–121] may not be compatible with durable construction.

The fewest possible different materials, which in turn should belong to distinct classes of materials, should be used, so that they can be more reliably identified for sorting after product use and the cost of recycling in terms of resources is lowered and the yield is improved [122]. Substances whose structure does not survive the recycling process (e.g., fibers, plastics, and organic working materials) have only limited recyclability.

Where it is not expedient to separate substances for recycling, they should be chosen such that materials with well-defined properties are obtained on melting, which can be substituted for materials manufactured from primary feedstocks (waste material compatibility).

Recyclability is enhanced when few composite materials are employed. However, composites, including materials with surface coatings, are important for functional performance and increased service life.

Joints and fastenings should be made so that they can be taken apart easily and nondestructively after product use [123], [124].

A prerequisite for recycling is the unambiguous identification of materials to enable their sorting and recovery as materials with well-defined properties. In this way, separation and sorting can make a formerly unrecyclable product recyclable.

Complex products should be designed in modular fashion so that the only part that has to be removed is the nonfunctional one. Another aspect of design for recycling is to make service and maintenance of products easy and cheap.

6.5. Integrated Measures to Reduce Environmental Burdens

Global optimization of environmental protection by integrated practices throughout the life cycle or for a plant or plant complex depends on a knowledge of the capabilities and limitations of the actions discussed in Section 6.4. Within the boundaries of a given system, an integrated, systematic analysis can help to harmonize specific actions to reduce burdens, sometimes with a departure from the achievable minimum.

6.5.1. Product-Integrated Environmental Protection

Environmental burdens due to a product are not fixed; they can be crucially influenced in the product design and development process. The cumulative ecological burdens of the four life cycle segments depend on

1) The materials (substances) chosen
2) The processes and technology used
3) The design and manufacture of the product
4) Product-related actions taken over the full product life cycle

All product-related actions must be taken so that all product properties specified in the requirement profile are guaranteed over the intended lifespan. The avoidance of product defects is the basis for the first three elements of product-integrated environmental protection (listed above) and is presumed in the following discussion.

The selection of materials governs the nature of the resources to be extracted. This fact results indirectly (through availability and renewability) in part of the environmental burdens due to the material. The choice of resources also governs the quantity of wastes and emissions as well as their pollution potential. The service life and the recyclability of the product depend on the materials and the design but also on the technology employed. The useful life of a product has a decisive effect on the magnitude of use-specific environmental burdens (Eq. 6.3) of the product, while the recyclability determines the cumulative ecological burdens of recycling EB_R.

The process can be optimized to increase the yield or efficiency by means of process control technology. Prerequisites for this include a means for the continuous measurement of process parameters in liquid or gaseous media. For example, the partial pressure of oxygen is measured with specially developed materials such as doped and stabilized zirconium oxide [125]; newly devised metallic and ceramic protective tubes for temperature measurement in melts or off-gas streams are able to withstand a wide range of ambient conditions over the needed span of time. Shortening of the process chain also depends on the availability of suitable materials. One example is the production of sheet steel with an as- cast thickness of 1 – 3 mm [115] by the process of near-net-shape casting [126]. Until recently, this process was impracticable because there were no materials for delivering the melt, no materials for protective tubes, and no sensors for continuous measurement of temperature and oxygen potential. Thin-slab casting by the SMS process with upright chill mold for the casting of slabs 40 – 50 mm thick has been in industrial use since the late 1980s [127]. A special chill mold with a ceramic dip tube for melt delivery has been designed for this purpose.

The choice of cooling method for turbine blades (base, convection, film, or transpiration) depends on the material (iron-base alloy, ceramic), the design, and the fabrication method. The design of the turbine buckets governs not only blade cooling but also the maximum gas inlet temperature and thus the energy efficiency of the turbine [112]. Increasing the energy efficiency leads to a drop in specific CO_2 emissions.

Suitable technology is needed to make materials with specific qualities. This dependence begins with the fabrication of strain-hardened metals and is especially vital in the manufacture of metallic glasses (material with longer life span), which are made at a cooling rate of 10^5 to 10^7 K/s. Mirror glass is now made by the float process on a molten tin bath and requires no subsequent grinding and polishing (shortened process chain). To produce deep-drawing steel without costly final processing, the oxygen converter is essential. Powder-metallurgical techniques have major advantages over machining in that the piece can be made with nearly its final dimensions, so that there is far less machining and material wastage [128].

Apart from optimizing the reduction of environmental impacts by decreasing the cumulative

ecological burdens, it is necessary to determine which conditions in the requirement profile prevent the use of more environmentally compatible materials and whether the profile itself could be modified. Changes in the profile and possibly in design, if they do not appreciably bear on the satisfaction of demand, make it possible to use materials and processes with smaller cumulative ecological burdens.

6.5.2. Production-Integrated Environmental Protection

The goal of production-integrated environmental protection in the chemical industry is to achieve an optimum, from the process engineering standpoint, in the process or the manufacturing complex [103]. A manufacturing complex consists of several processes, which need not be directly linked. The term production-integrated environmental protection was developed in the chemical industry; today it is used in contrast to product-integrated environmental protection, which refers to optimization of environmental burdens reductions within a company or production plants. In production-integrated environmental protection, all chemical, physical, biological, and process and manufacturing engineering possibilities are utilized. The scope of the resulting process improvements can range from the optimization of single process components and extends up to modification of entire processes or of manufacturing complexes. It includes:

1) Primary waste prevention in the process.
2) Secondary waste prevention, that is, recycling of wastes in the manufacturing complex so that they can be used as feeds to other processes, or for energy production as the last and completing stage in the environmentally sound disposal of substances that cannot be reclaimed (waste for disposal, wastewater, and off-gas).

While additive environmental protection always leads to extra costs, the financial burden on the company can be reduced when production-integrated environmental protection is practiced.

Integration of two fundamentally different pyrometallurgical reactions in one vessel (QSL technology) made it possible to optimize the conventional process for lead refining, which consists of a traveling-grate sintering machine, in which the concentrate is roasted and sintered, and the blast furnace, in which the sinter is remelted under reducing conditions. In the cylindrical QSL reactor, autogenous roasting reaction melts of sulfur- and lead- containing feeds is followed by carbothermic reduction of the lead oxide bound in the slag. Along with the positive energy balance, emissions were cut relative to the conventional process: arsenic from 278 to 72 kg/a, lead from ca. 2000 to 150 kg/a, cadmium from 250 to 16 kg/a, and sulfur dioxide from 5230 to 89 t/a [129].

Improvements in chemical processes by using new synthesis routes and new combinations of process steps made it possible to develop low-waste processes in areas such as the high-tonnage production of polyethylene. In 1964, losses to the atmosphere were 9 %; to wastewater, 6.4 %; and to landfill, 1.3 %. With the new process, losses in 1986 were just 1.6 % in the boiler house and 0.4 % in combustion and landfill disposal of residues [130]. In the production of diphenyl bases, reduction with zinc has been replaced by a catalytic process. Solvents and sulfur acids are regenerated and recycled. So the generation of zinc oxide (formerly about 8000 t/a in Germany) is avoided [131], [132]. In polyester synthesis, process optimization in terms of plant and control systems made it possible to integrate recycle streams and purification steps into the plant so that virtually the only substances leaving the plant apart from the product are CO_2 and CO [130].

Gaseous, liquid, or solid wastes generated in a process, even under optimal operating conditions, can often be forwarded to several processes in the manufacturing complex. An example is the reclamation of residues by recycling directly to the process, for example return slag in the blast-furnace process [133], together with the use of such residues as feeds for other products.

An example of internal recycling in a metal surface-treatment operation is shown in Figure 30. By eliminating the introduction of foreign substances and electrolyte loss by entrainment, the number of rinse stages can be reduced, and the quantity of waste separated in the regenerator decreased. Reduced electrolyte entrainment also diminishes the volume flow rate $\dot{V}_R$

to the concentrator. This may even make it unnecessary to use a separate concentrator (e.g., an evaporator). The process liquor flow rate $\dot{V}_P$ is directly proportional to the inlet rate of foreign substances (e.g., organic contaminants in electrolytes, iron in hydrochloric acid pickling liquors, copper in peroxodisulfate etching liquors, oils in alkaline cleaning solutions). If this inlet rate of foreign substances can be reduced, the volume flow rate of process liquor $\dot{V}_P$ can be decreased, and this may permit the use of a smaller regenerator (ion exchanger, membrane electrolysis, ultrafiltration) [134]. Actions to reduce environmental burdens must be matched to the metal surface treatment process and may lead to dramatic changes in effluent relative to the open electrolytic system: Wastewater production cut by around 98 %; pollutant burden in wastewater lowered by 99 %; solid wastes, especially cell slimes, reduced by ca. 95 %. The concept illustrated has been used in a number of zero-wastewater electrolysis operations [135].

Changes in process and apparatus used for painting can increase the application efficiency and thus reduce the amount of paint sludge (see Section 8.4). Possibilities in spray painting include electrostatic spraying and airless spraying [136]. If the paint system is changed from solvent-based to water-based, solvent emissions can be markedly lowered. Recycling in the spray painting operation can lead to a 60 – 70 % reduction in paint-sludge wastes. The prerequisite for recycling paint in marketable quality is that overspray (paint which has missed the workpiece) be intercepted with no change in quality. A number of closed circuits are used. In a first cycle, paint is recovered in the paint booth by setting plates and belts behind the workpiece to catch part of the overspray and return it directly to the paint loop. A second cycle recovers paint not picked up in the first, which is discharged as paint sludge from off-gas cleaning. The sludge is dewatered, solvents are added, the liquor is sieved or centrifuged and depigmented, and paint is recovered. If the quality is to be maintained, unsuitable coagulants must not be introduced in the scrubbing step. The loop is closed by a third cycle in which the solvent escaping from the spray booth is recovered. This can be performed by adsorption on activated carbon followed by desorption of solvents, which can be recycled to the paint or used in cleaning the painting system [137], [138].

6.6. Determination of Environmental Burdens

The basis for optimizing environmental burdens is an objective determination of cumulative ecological burdens; that is, the sum of the environmental burdens resulting from the system under consideration. The instrument used is the "life-cycle assessment" method now being standardized (see Chap. 4) [105], [139]. This is a holistic approach to the objective and verifiable determination of environmental burdens caused by the object under study (product, company, plant, service, disposal route). The life- cycle assessment method is broken down into the following stages: goal definition and scope, inventory, impact assessment, evaluation, and improvement assessment (optional). Any life- cycle assessment must be presented in a comprehensible and verifiable way and must take account of the product's useful life. Accordingly, the goal definition must begin by stating the study purpose. Life-cycle assessment can be used for a variety of purposes: to optimize the object under consideration, to select among alternatives (products, materials, processes), or to develop environmentally benign products. The goal definition states, among other things, what object is being studied (product, company, service, or disposal route); what quantity will be used in the balance calculation (the functional unit); the time period of considered data; and the geographic and methodological boundaries of the systems compared.

In the inventory, material and energy inputs and outputs are balanced for each process within the system boundaries and aggregated according to the contributions of process products to the functional unit. The impact assessment aims to determine the potential impacts of the material and energy balances from the inventory in the most realistic way possible. The material and energy balance data are determined under impact categories such as resource scarcity, greenhouse effect, and damage to the ozone layer, and with equivalency factors such as years, global warming potential, and ozone depletion potential. At the current state of the art, the impact assessment can only be done for partial areas. The last part

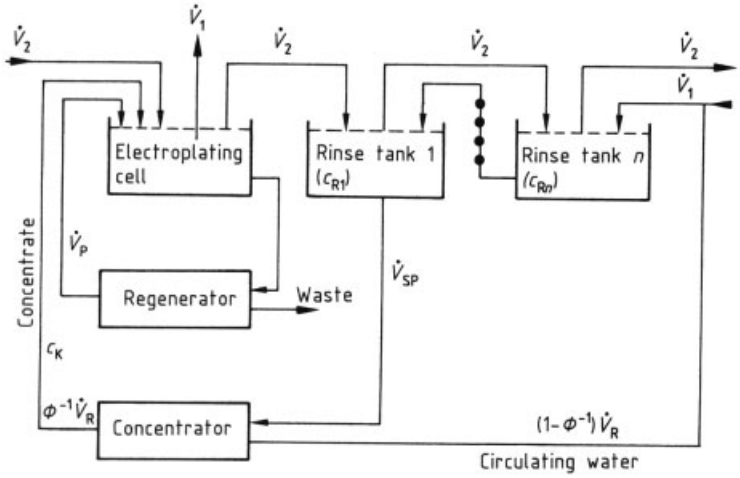

Figure 30. Example of recycle streams eliminating wastewater generation in an electrolytic coating plant [134]
Φ is the degree of concentration; V_1 is the volume flow rate evaporated; V_2 is the rate of electrolyte entrainment; V_R the rate of rinse water; V_P the rate of process liquor; and c the concentration

of the life-cycle assessment, evaluation, is the least well developed at present [140]. Despite these limitations, if the inventory is carried out by a sound method and the impact assessment is taken as far as is now possible, the results can be a very good decision aid. Far-reaching and complex relationships can be reduced to a few easily surveyed data sets. Thus life-cycle assessment provides the only objective basis for reducing environmental burdens by optimization, selection, or development. Life-cycle assessment in practice is an instrument for measuring environmental burdens and reductions in burdens brought about by the system under examination. Practices of integrated technical environmental protection, like those of technical waste prevention, can now be consistently quantified with the help of life-cycle assessment.

7. Recycling of Waste

7.1. Introduction

7.1.1. General Aspects

The waste management includes recycling and various disposal operations. In waste recycling, the main purpose is to utilize the waste, not to eliminate its pollution potential. Disposal operations include physical, chemical, and thermal processing, as well as placement in landfills. According to the German Recycling Act (Kreislaufwirtschaftsgesetz) [141], wastes for recycling are thus distinguished from wastes for disposal.

Recycling is the measure of the secondary part of waste prevention (see Chapter 6.1) and includes not only reclamation but also the re-use of products or parts of products. Recycling means returning production wastes, post-consumption wastes, products, and parts of products to the production/consumption loop with no loss of quality. The term "recycling" applies only when the secondary raw material or secondary product generated can act as a substitute for a substance or a product of primary production. The term "upcycling" is used when an improvement in quality is achieved through recycling, and "downcycling" when the process is associated with a loss of quality.

In every form of recycling, the objectives are to preserve primary resources and landfill capacity, which is equivalent to reducing environmental burdens by diminishing the amount of wastes and emissions produced.

7.1.2. Forms of Recycling

A distinction is made between recycling of materials and recovery of energy content [142]. Recycling of materials is subdivided into various

forms of use and reclamation, which are shown schematically in Figure 31 along with consumption (use) and production. Each of the nine activities shown in the figure (numbered boxes) consumes primary resources and generates wastes for recycling (from boxes 3 – 7 to box 8) or for processing (from boxes 1 – 8 to box 9). The secondary resources (SR) generated in each activity become feeds for production. The term "waste" includes used products, used components, used assemblies, waste materials, and residues.

Recycling of Production Waste. Production wastes (Fig. 31, box 3) are all solid and liquid products that do not meet process objectives (product, coupled product), and are not wastewater. Production waste recycling is the return of production wastes to the same or a different production process. Examples of production wastes are sprues from injection molding, trimmings and stamping wastes, overspray from spray painting, and used lubricants. A characteristic of production waste recycling is the dissolution of the shape of solid wastes. This form of recycling is highly developed in some cases because wastes are collected in pure form and the knowledge required to process them is available. Often, production wastes can be processed in the same plant and returned to primary production.

Recycling during product use (reuse) (Fig. 31, box 4) means returning used products, parts, and assemblies to consumption without dissolution of their shape. Here "product" means goods for use and for consumption, equipment, and infrastructures (buildings and transportation and communications systems). Used products may be disassembled (nondestructively), cleaned, inspected, reconditioned or reworked, and reassembled. Examples include industrial robots, replacement auto engines, and photocopiers. Recycling during product use is characterized by the preservation of form and also by the use of manufacturing processes for reconditioning or reworking. In all other forms of recycling, wastes are processed in some way. The re-use of processed products for the same use, or the utilization of remanufactured products for a purpose different from the last means high-level recycling. As a rule, the consumption of primary resources is less than in other forms of recycling and the environmental burdens are therefore slighter.

Recycling of used products (Fig. 31, boxes 5 to 7) falls into three classes: product-specific recycling (box 5), material-specific recycling (box 6), and recycling of feedstocks (box 7). Which applies to a given used product depends on the degree of complexity, the construction, the composition, and the quantity. Recycling of used products typically involves breaking down the shape of the used product.

Product-specific recycling (Fig. 31, box 5) means returning the materials making up the used product as well as process media to production, following a procedure adapted to the product. Complex products containing many parts are processed by specially developed methods. Examples are automobiles, large household appliances, and computers. Still serviceable parts or assemblies are withdrawn in partial disassembly of the used product and forwarded to recycling during product use. Substances that can be sent to material-specific recycling and those that would interfere with further processing (copper, tin from steel bodywork) are also removed. In mechanical separation, the complete or partially disassembled used product is ground [143]. This is followed by sorting into fractions based on differences in physical properties (shape, density, particle size, electrical and magnetic properties, etc.) or possibly in chemical composition [144]. The various separation stages in product-specific recycling yield usable parts and assemblies, secondary materials for production, materials for material recycling and/or substances for recycling of feedstocks, and wastes for recycling or treatment. Components containing more than one substance may be passed on without further mechanical processing to thermal processing; for example, lead batteries may go to blast-furnace processing [145], or partially disassembled products such as compacted auto bodies may go to "metallurgical recycling" [146]. As a rule, product-specific recycling is accomplished by the method that entails the least consumption of primary resources compared to other forms of recycling of used products.

Material-specific recycling (Fig. 31, box 6) refers to obtaining well-defined materials from one or several used materials. In this context

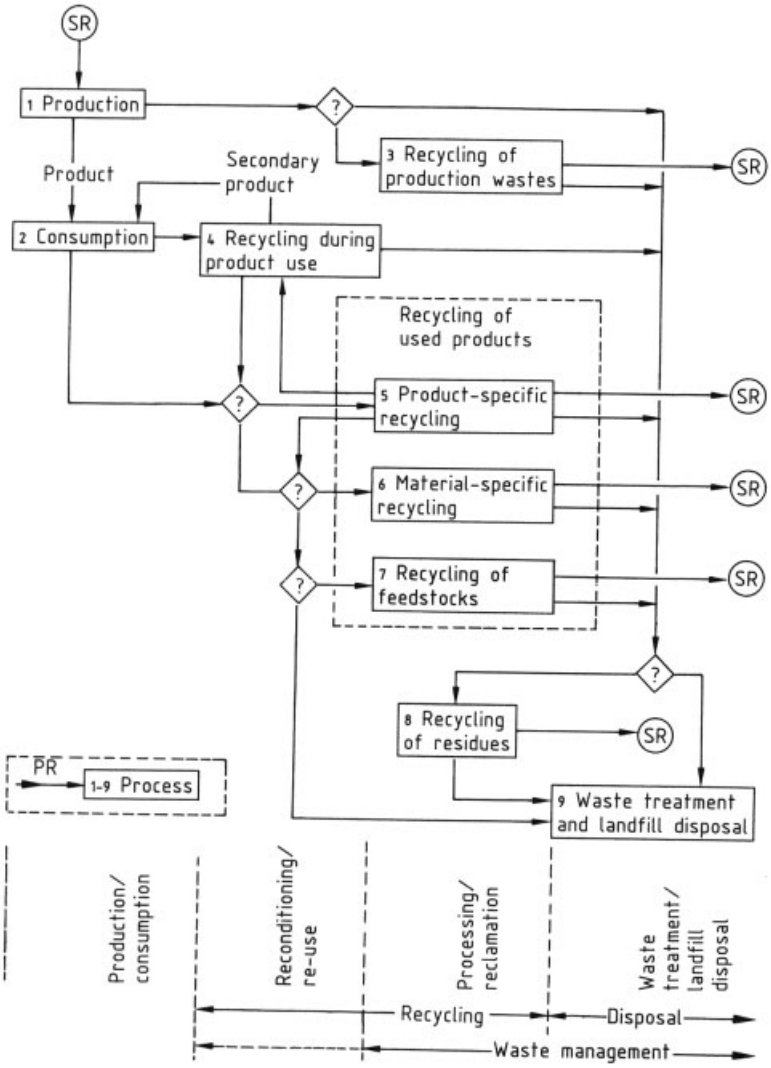

Figure 31. Forms of recycling and their links to production, consumption, and waste disposal
PR denotes primary resources; SR, secondary raw materials

the notation of "material" is extended to include liquids and process media. These are fed into processes to make semifinished or finished products. Characteristic of material-specific recycling is that only physical operations are employed; there is no chemical transformation of the used substances. After sorting and washing, used metals and plastics are generally melted down to obtain well-defined materials. Because material-specific recycling does not involve chemical reactions, the inlet materials must be pure enough not to interfere with the production of a well-defined material (i.e., they must contain only small amounts of impurities, contaminants, and incompatible substances). Substances that are remelted together must be compatible in the sense that they can be made into a well-defined material, possibly after compensating for loss of alloying elements and auxiliaries. The prerequisite for this is collection in pure form. In the case of used plastics, material-specific recycling is referred to as mechanical recycling.

Recycling of feedstocks (Fig. 31, box 7) means the use of chemical and physical processes to return used substances to the production process. In contrast to material-specific recycling, here classes of substances together with foreign substances (deposits or impurities) can be used. The prerequisite, however, is that substances interfering with recycling can be separated by chemical reactions if physical separation is not possible. Typical of this form of recycling is that at least one separation step involves chemical reactions. The products of these processes that serve process objectives are fed to production processes and used in producing materials. In the case of used plastics, this form of recycling is also called chemical recycling.

Recycling of Residues. In residue recycling (Fig. 31, box 8) residues from a recycling processes (boxes 3 – 7) together with production residues are reclaimed. Such residues include, for example, slags, drosses, dusts, and sludges. Recycling of residues is characterized by a change in the nature of the substance in chemical reactions. The residues, after preliminary processing when needed, are fed into primary production processes for two reasons: These processes are suitable for the recovery of valuable substances, and no recycling technology is available for certain wastes. Residues from production forwarded to residue recycling differ from recycling process residues in being pure (sorted) and free of foreign substances.

Recycling for Energy Recovery. According to the German Recycling Act, wastes can be recycled as materials or used to produce energy. The method selected will be the one that causes the slighter environmental burdens. The term recovery of energy content is used when wastes are used as a replacement fuel and the principal aim is to utilize the heating value, not to reduce the pollution potential [141]. Thermal processing to dispose of wastes is not the same as recycling for energy recovery. The Recycling Act permits energy recovery only when a particular waste, unmixed with other substances, meets the following tests:

1) The heating value of the waste (except for wastes from renewable resources) is at least 11 000 kJ/kg
2) A furnace efficiency of at least 75 % is attained
3) The heat produced is utilized
4) Wastes generated by the process can be placed in landfills with the minimum possible further treatment

7.1.3. Prerequisites for and Limitations of Recycling

Recycling as materials can be said to occur only if it is possible to substitute secondary resources for primary resources in production. Recycling for energy recovery must replace primary energy resources. While there are quality changes in upcycling and downcycling, these two special forms of recycling must meet the same criteria as recycling of materials.

The basic precondition for recycling is collection of production wastes, used products, and residues. The form of collection and the logistics involved in delivering the materials to the recycling plants have significant effects on the consumption of primary resources and thus on the environmental burdens, the costs, and the economics and quality of the secondary products or secondary materials. Primary resources can be replaced by secondary materials only if the product quality is maintained. Together with a lack of universal acceptance for products made from secondary materials, other constraints arise from issues of product liability, technical codes, legal restrictions, unrealistic demands from consumers, and excessively strict requirement profiles written into specifications. The basis for overcoming these hindrances is consistent, guaranteed quality of recycled products [147].

In addition, it is essential to make certain that recycling actually does relieve environmental burdens in comparison with the use of primary resources and that recycling is economically feasible. These prerequisites can be met only if the following conditions hold:

1) Adequate recycling logistics (collection and transportation)
2) Sufficient quantity of suitable waste (production waste, used products, waste substances, and residues) to keep the recycling facility utilized over its depreciation life

3) Adequate demand for secondary products and secondary raw materials over the depreciation period
4) Availability of sorted wastes without interfering deposits or impurities
5) Recyclability of the used products (design for recycling, recycling compatible materials, etc.)
6) Dependable means of disposal (treatment and/or landfill deposition) of wastes generated in recycling
7) Availability of suitable recycling technology

The limitations of recycling can be classed as ecological, economic, and technical. To decide whether recycling is ecologically or economically desirable, it is compared with disposal of the waste in question. Recycling is economically preferable to disposal if the sum of the proceeds from secondary feedstocks and non-recycling disposal costs is greater than the cost of recycling (Eq. 7.1). The point at which proceeds equals the difference in recycling and disposal costs is the breakeven point [148].

$$P_{\mathrm{SR}} + C_{\mathrm{D}} > C_{\mathrm{R}} \rightarrow P_{\mathrm{SR}} > C_{\mathrm{R}} - C_{\mathrm{D}} \tag{7.1}$$

$$M_{\mathrm{SR}} > \mathrm{EB_R} - \mathrm{EB_D} \tag{7.2}$$

where C stands for cost, P for proceeds, M for the ecologically weighted mass, EB for environmental burdens, D for disposal, SR for secondary resources, and R for recycling.

The ecological limit of recycling occurs at the ecological breakeven point (ebep) [149]. Recycling is preferable to disposal for a given waste only when the decrease in environmental burden, represented in Equation (7.2) by the ecologically weighted mass [149] of the secondary raw material, is greater than the difference in environmental burdens between recycling and disposal (Eq. 7.2). The environmental burdens in each case are the sum of all resources used and all wastes and emissions generated within the system under analysis. Environmental burdens are determined by the life cycle assessment method. The secondary raw material produced by recycling is not in itself a quantifiable reduction in environmental burdens. Because this secondary resources replaces a substance that would otherwise have to be made from primary resources, the decrease in environmental burdens is equal to the burdens that would arise in the production of this secondary feed from primary resources. These theoretical burdens make it possible to quantify the decrease in burdens as M_{SR} in Equation (7.2). The ecological breakeven point occurs when the left- and right-hand sides of this equation are equal. If the ecological result of recycling is below the breakeven point in this comparative analysis, the contemplated form of disposal is preferable to recycling from an ecological standpoint.

Technical limitations on recycling arise from the fact that logistical systems and recycling methods are not available for some wastes and the attainable quality of secondary feeds or products may not be sufficiently high. Recycling for recovery of materials also faces a technical limitation when the recycling technology damages the substances (e.g., paper and plastics).

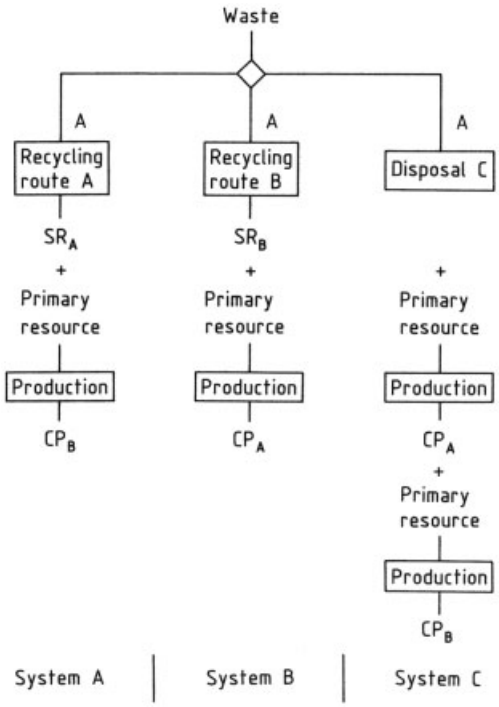

Figure 32. Three options for the handling of a waste
The utilities of the three systems are made equal by adding "complementary goods" CP_x to the secondary raw materials SR_y

7.1.4. Selection of Disposal Route

Figure 32 shows alternative routes for recycling as materials and waste treatment (waste processing and landfill disposal). A third option is energy recycling. For a given waste, the option chosen must be technically feasible and economically reasonable and cause the smallest total environmental burdens.

The precondition of recycling is technical capability. The economic aspect is based on a pure economic analysis. The ecological considerations involved in this choice are more complex, and it is these that are discussed in what follows.

The life cycle assessment technique (see Chap. 4) is used to determine environmental burdens for each pathway. In this partial assessment for the post-consumer sector, it is not possible to determine a functional unit. The waste at the beginning of each route is identical in quality and quantity; the quality and quantity of the secondary resources are a function of the particular recycling or disposal option. The benefit of a commodity, in this case a secondary resource, comprises objective and subjective factors, i.e., measurable and unmeasurable quantities. As a result, each pathway generates secondary resources differing in benefit, and this benefit is generally not measurable. If the options are nevertheless to be compared, the benefits of the systems must be made equal. The way of doing this is to render the goods produced in each system identical in quality and quantity. This is done, as shown in Figure 32, by placing complementary products (CP) in the "shoppingbasket" of each system. When two options are being compared, the complementary products added to one correspond, in quality and quantity, to the secondary resources produced in the other. Because the complementary goods generated from primary resources correspond in quality and quantity to the secondary resources, the baskets are equal in benefits and the systems can be compared (see Eq. 7.3) [150].

$$\mathrm{SR}_x = \mathrm{CP}_x \tag{7.3}$$

$$\mathrm{EB}_\mathrm{A}^\mathrm{S} + \mathrm{EB}_\mathrm{B}^\mathrm{P} \Leftrightarrow \mathrm{EB}_\mathrm{A}^\mathrm{P} + \mathrm{EB}_\mathrm{B}^\mathrm{S} \tag{7.4}$$

where SR denotes secondary resources; CP stands for complementary products; EB are environmental burdens, A and B are goods; and P and S refer to goods produced from primary and secondary resources respectively.

It is not the benefits that are compared; these are equal. The comparison is between the environmental burdens of the two equal-benefit systems (Eq. 7.4). The option with the smaller environmental burdens is preferable from the ecological standpoint for the given waste.

7.1.5. Recycling Rates

Recycling rates are calculated in different ways and may therefore be reported differently for a given situation. The recycling rate should reflect what fraction of the material in question (from production process 1) is contained, after use and recycling, in the resulting products from process 2. The relationship is described by equation (7.5) for one substance; Figure 33 illustrates the operations and the notation. The equation holds for materials or substances derived from used (waste) products and residues.

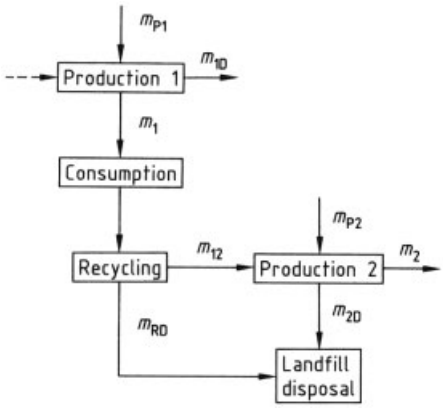

Figure 33. Block diagram used in deriving equations for the recycling rate

The equation does not apply to the recycling of production waste but can be extended to this case; it is therefore usable for both closed-loop and open-loop recycling. Equation (7.5), and also Equations (7.6) – (7.8) derived therefrom, presuppose that the masses under consideration have been adjusted for changes in stock as well as exports and imports. If the losses in the second production process (mass flow rate $m_{2\mathrm{D}}$) are zero, Equation (7.5) becomes Equation (7.6). The recycling rate is the ratio of the

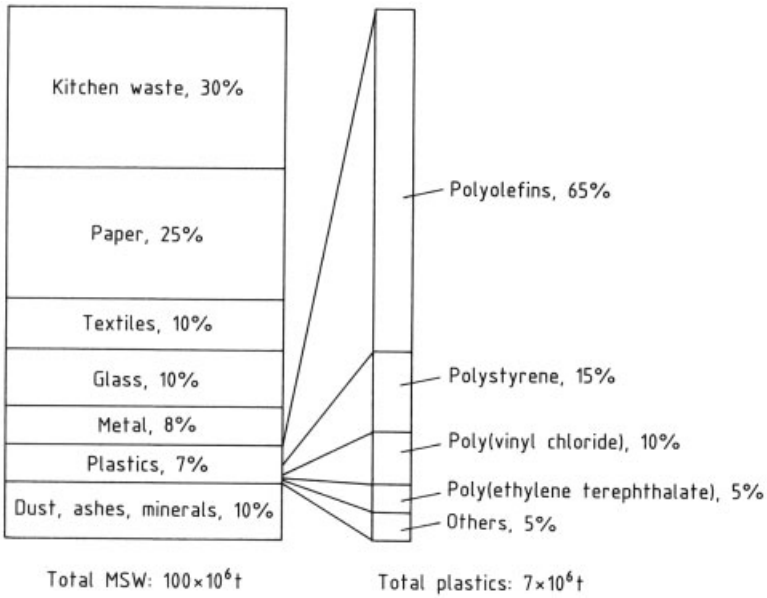

Figure 34. Quantitative breakdown of wastes in Western Europe

mass flow m_{12} into the second process to the rate m_1 into products from the first process. These two ways of determining the recycling rate are seldom used because it is difficult to relate the argument to m_1, especially when long spans of time intervene between the first and second production processes, as is the case with durable goods.

$$R_{\mathrm{C}} = \frac{m_{12}}{m_{12} + m_{\mathrm{p2}}} \cdot \frac{m_2}{m_1} \tag{7.5}$$

If $m_{2\mathrm{D}} = 0$, then $m_2 = m_{12} + m_{\mathrm{p2}}$ and

$$R'_{\mathrm{C}} = \frac{m_{12}}{m_1} \tag{7.6}$$

If $m_1 = m_2$,

$$R_1 = \frac{m_{12}}{m_{12} + m_{\mathrm{p2}}} \tag{7.7}$$

If $m_{2\mathrm{D}} = 0$ and $m_1 = m_2$,

$$R_2 = \frac{m_{12}}{m_2} \tag{7.8}$$

where R_{C} is the "correct" recycling rate (the other R being the recycling rate derived from this), and m_{x} are the stream volumes as shown in Figure 33.

If m_1 is set equal to m_2, it is not necessary to consider two time intervals, and Equation (7.5) becomes Equation (7.7). This form of the recycling rate is frequently chosen; the error is larger, the more m_1 and m_2 differ. If the assumptions used to derive Equations (7.6) and (7.7) are applied together, Equation (7.5) takes the form of Equation (7.8). Here the recycling rate is the simple ratio between the mass flow rate m_{12} inlet with secondary resources to the mass flow rate m_2 discharged. This figure is also often quoted as the recycling rate. Of all the formulas listed, it has the largest potential error.

For a given situation, the recycling rate can vary widely depending on what relation is selected. In neither case, however, does it say anything about whether recycling is ecologically desirable in the given case. Firstly, the analysis is generally based on averaging all recycling results for a given substance within a time interval; secondly, no information is given here about the environmental burdens due to recycling or the reductions in burdens that result.

7.2. Plastics

7.2.1. Introduction

In all densely populated areas of the industrialized countries, recent years have seen major

bottlenecks develop in waste disposal, because of exhaustion of landfill capacity as well as a lack of acceptance of waste combustion plants. Legislators in some of the hardest-hit countries have been driven to impose harsh restrictions. Some of the new provisions discriminate against plastics in general or certain types of plastics.

Examples are found in acts and regulations issued in Germany, which are based on two key principles:

1) The polluter pays principle, which transfers responsibility from communities to parties generating waste or marketing products that will become wastes. Examples are the Packaging Regulation (Verpackungsverordnung) of 1990 and the Regulation on the Disposal of Used Autos (Verordnung über die Entsorgung von Altautos).
2) The recycling principle, which is embodied in the new Federal Waste Act (Abfallgesetz) and states that reuse has priority over other disposal practices for all products.

The Packaging Regulation, already in force in Germany, not only confuses the total waste streams for paper and glass and discriminates against plastics as a packaging material, but also imposes vastly increased prices for classical disposal routes, in particular landfill disposal. Costs for plastic wastes rated as nontoxic are currently at least 260 €/t – 100 times more expensive than 10 years ago.

A provision that merits especially critical examination is the ban on incineration and thermal utilization of waste plastics packaging. No such provision can be found in regulations now in effect or contemplated for the rest of Europe, and even in Germany this act may be subject to revision [151]. The second plastics-related problem with the German regulation is that its deadlines are too short and its quotas too high. These have already brought *Duales System Deutschland*, a new concept set up to collect household packaging wastes, to the brink of ruin.

Household waste and its disposal is the most serious problem in the case of plastics, because this type of waste is highly soiled and almost inseparably mixed with other plastics, paper, and metals. Figure 34 shows a quantitative breakdown of wastes for the EU countries. Plastics represent only 7 wt % of residential waste and thus represent a minor problem in the total waste streams, 70 % of which currently is destined for landfills. This is so in spite of the fact that the volume fraction of plastics in a landfill may be a factor of 2 – 3 greater than the weight fraction.

Current widespread public aversion to the chemical industry, and hence to plastics, is hindering approval of specialized disposal methods. Plastics recycling has therefore been exemplified by mechanical recycling – long regarded by experts as suitable for only a small fraction of plastics wastes – as is typical for competing materials such as metals and glass.

7.2.2. Mechanical Recycling

7.2.2.1. Chemical Basis for Plastics Recycling

At a certain elevated temperature, determined by the polymer chain structure and chain length, thermoplastics soften and melt because the molecules can move independently of one another. This movement intensifies as the temperature rises. In this state, the molecular structure can be thought of as a "felt" of threads, with individual molecules forming coils whose outer ends are entangled with one another (Fig. 35 A).

Polymers do not exhibit a vapor state because the macromolecules are cleaved to hydrocarbon fragments before they can vaporize.

Polymer molecules can also be degraded at low temperatures, for example as a result of severe mechanical stress, the action of moisture and chemicals, or radiation, such as ultraviolet rays. This kind of damage, known as aging, chiefly affects molecular chains extending from the outside of the coils; these are the portions of the molecules that formed the entanglements (Fig. 35 B). A degraded coil is thus linked to its neighbors by only a few entanglements, so that the melt is less viscous. This is often an advantage in processing, but it also reduces the toughness of the solid plastic, leading to impact brittleness.

Polymers with different molecular structures repel one another on solidifaction. Thus a melt of mixed plastics segregates on cooling, forming phases that no longer cohere with one another (Fig. 35 C). The second polymer forms a separate phase, which encloses the coil of the first

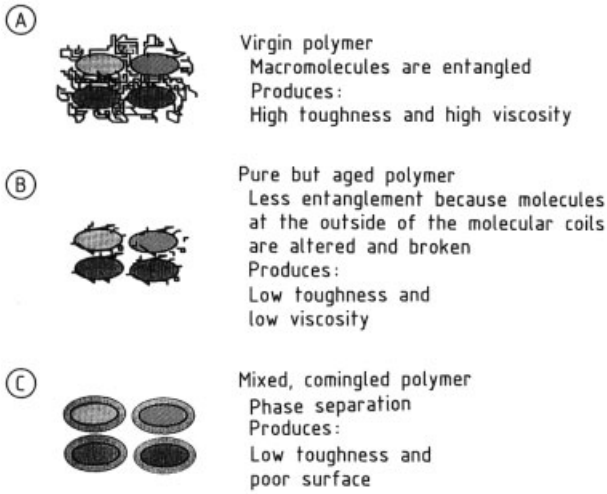

Figure 35. Changes in the physical structure of thermoplastics

resin like a shell, forming an interface that interferes with the entanglement of like molecules. Components made of such polymer blends have adequate mechanical properties only if the second phase is uniformly distributed and very finely dispersed in the first, or if chemical additives can be used to keep the unlike molecules from repelling one another and allow them to become entangled.

In mixtures of polyethylene with up to 10 % polypropylene (or vice versa), such fine dispersion can be achieved with compounding screws, and good properties can be obtained. When the phases are too large, they lead to surface structure variations, apparent as optical defects.

7.2.2.2. Pure Wastes

Broad experience has been gained with recycling pure thermoplastics. All manufacturing plants have processed pure wastes by grinding and incorporating them into new products since the beginning of thermoplastics production. A typical modern processing machine that can incorporate pure, clean recycled material into virgin stock is an injection-molding machine in which a bypass extruder melts recycled plastic granules and meters the melt into virgin stock through a fine filter. Up to 30 vol % of recycled plastic can be introduced.

In the automobile industry, the removal of plastic components from used bodywork and their remelting to make new moldings has already become standard practice [152–154]. Automakers now accept products made from recycled plastics or containing a fraction of recycled plastic, provided the recycled granules meet certain specifications of uniformity and quality, certified by neutral testing [155].

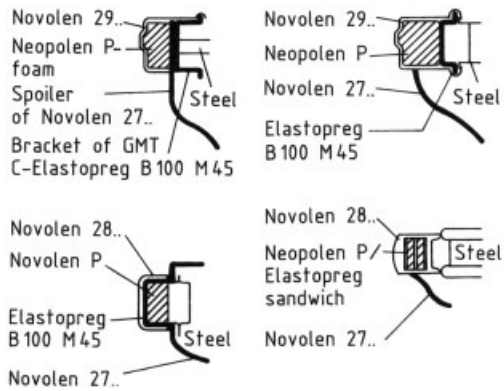

Figure 36. Polypropylene bumpers (BASF)

The requirement to use a higher proportion of recyclable plastic components in consumer products, including automobiles, has already led to redesign. The variety of materials has been narrowed, and composite parts have been replaced by assemblies made of a single material. As a result, simple disassembly yields large amounts of pure material for recycling. A typ-

ical example of a good design of this type is the bumper assembly shown in Figure 36, in which all the components are made of polypropylene. The shock-absorbing part is made from a newly developed polypropylene foamed-pellet material [156], and the supporting bracket from polypropylene with long-fiber glass reinforcement. This construction makes it possible to remove 30 % of the plastics in a modern automobile in about 30 min [157].

7.2.2.3. Mixed and Soiled Plastics Wastes

Products from Mixed and Contaminated Plastics Wastes. There are several processes by which useful objects, such as planks and posts for park benches or brick substitutes, can be produced from mixed and unpurified resin wastes. Figure 37 shows a plant consisting of pelletizing and compounding units as well as plasticizing and compression-molding machines. The market for such articles is limited, and they are difficult to sell because the items they replace are cheap. Up to now, only subsidized products have survived on the market.

Separation of Mixed Plastics Wastes. To make competitive products of adequate quality from contaminated and mixed plastics wastes, separation into pure fractions is necessary. The steps required are as follows:

1) Grinding to pellets ca. 1 mm in diameter (an expensive operation)
2) Washing
3) Separation of particles by density (at most two different plastics with ca. 99 % purity per operation, so economical separation depends on minimizing the number of plastics in the mixture)
4) Drying

Experience in recent years has shown that this method of processing into new plastics does not offer competitive prices and thus can be considered only if subsidized.

Dissolution. Thermoplastics can be dissolved in certain solvents. This technique used to be quite important in the production of films and fibers and is still used for some lacquers (e.g., nail polish) and adhesives for hobbyists. On environmental grounds, the technology is virtually extinct in industry. The solvents, most of which are toxic, could not be completely recovered from the products, and the equipment could not be sealed tightly enough to prevent them from escaping into the surroundings.

The dissolution of thermoplastics is based on the chemical affinity of some liquid hydrocarbons for certain resin molecules. When a suitable solvent comes in contact with a plastic, it diffuses into the material and penetrates the tangle of molecules. When enough solvent has been absorbed to fill the molecular coils, the plastic begins to swell. If more solvent is added, the molecules separate from one another to form a solution. This process is reversible, and if the solvent is evaporated, the polymer remains behind as a film.

If a selective solvent is used, only polymer molecules of one certain type are dissolved. It is even possible to treat mixtures such as thermoplastics in household waste with a single highly selective solvent, and by varying the temperature and pressure to dissolve several pure plastics one after another [178]. Dissolution is then a physicochemical method for recovering pure plastics from mixed plastics wastes. An example of the selective dissolution of mixed plastics wastes with THF follows:

First cut:	room temperature	
	Poly(vinyl chloride)	> 99 %
	Polystyrene	> 99 %
Second cut:	70 °C	
	low-density polyethylene	> 99 %
Third cut:	160 °C	
	polypropylene	> 99 %
	high-density polyethylene	> 99 %
Fourth cut:	190 °C	
	poly(ethylene terephthalate)	> 99 %

The dissolved fractions can be purified by filtration before the solvents are removed by evaporation.

A pilot plant with a capacity of 50 kg/h was started up in the United States. Costs for the recovered pellets are estimated at ca. 0.4 €/kg for a 25 000 t/a plant. Because of strict environmental regulations, this process would be far more expensive in Europe and therefore would be limited to special cases.

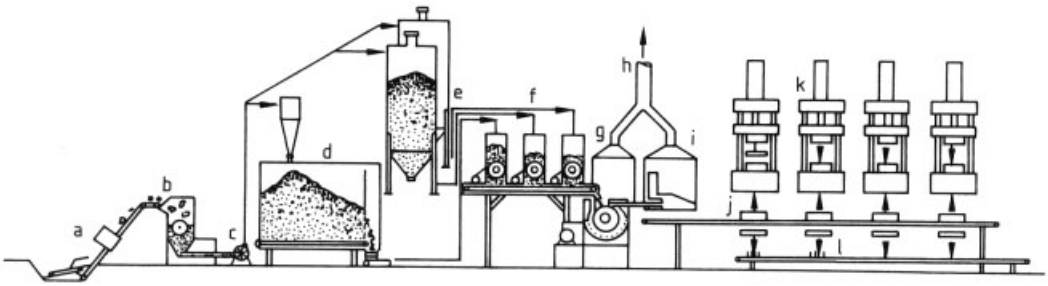

Figure 37. Fabrication of thick-section parts from mixed platics wastes (Meinberg)
a) Charging, separation of metals; b) Chopping mill; c) Suction conveying; d) Bunker storage; e) Mechanical conveying; f) Metering unit with star feeders; g) Plasticator (1000 kg/h); h) Degassing, gas scrubbing; i) Volume metering unit; j) Automatic press charging; k) Press (3000 kN); l) Automatic discharge

7.2.2.4. Fiber Composites and Other Thermosets

Thermosets are cross-linked resins that cannot be melted because the process that created the cross-links is irreversible. Typical products are auto body parts made from sheet molding compounds (SMC) which contain only 30 vol % of the resin together with 60 vol % fillers and reinforcing fibers. Pure wastes of this material can be pulverized in a hammer mill [158] and the powdered resin and fibers recovered, and used as additives for new moldings. The majority of the glass fibers are obtained in usable lengths, and are almost completely separated from the binding resin, which is reduced to dust. The mill discharge can easily be separated into three or more fractions:

1) Powder consisting of resin and short glass fibers (<0.5 mm)
2) Fiber bundles with fiber lengths up to ca. 5 mm
3) Fiber bundles with lengths > 5 mm

The powder fraction is added to a new batch of resin in amounts of up to 20 %, replacing mineral powder filler. Most of the middle fraction can also be incorporated in new resin. The long-fiber fraction cannot be utilized in SMC because it would interfere with the spreading of the resin paste. But it can be used as reinforcement in short-fiber glass-reinforced thermoplastics or in molding compounds. The prerequisite is an adequate and dependable supply of pure recycled material of uniform quality.

7.2.2.5. Outlook for Mixed Wastes in Blends

Modern blending technology could also be applied to mixed plastics wastes [159]. However, the blend must be completely homogeneous and must not change over a long time span. The wide variations in properties that would result from any fluctuation mean that blends are limited to a few applications. Because this technology will be restricted to special cases, it is not discussed further here; the reader is referred to [159].

7.2.3. Degradation to Molecular Constituents

In theory, the relative ease of degrading polymers into their molecular constituents should allow their disposal, or the monomers of low molecular mass polymers obtained by degradation could be repolymerized. Both methods are under investigation.

7.2.3.1. Biodegradable Plastics

Physically, polymers with their long-chain molecules represent a higher state of ordering than the small monomer building blocks. They therefore break down under the environmental conditions. Modern plastics, however, are designed for long service life and contain chemical stabilizers, so that they retain their useful properties over a long period under normal conditions.

The degradation process can be accelerated simply by omitting the stabilizers or even by

adding initiators that increase the rate of degradation. This approach leads to plastics that degrade spontaneously, also called biodegradable plastics.

In the ideal case the degraded plastics would break down into nontoxic, naturally occurring molecules. Such materials are commercially available. They disintegrate into readily degradable molecules derived from natural products, such as starches. Plastics of this kind are already in use on a limited scale. An example is Biopol, a biodegradable plastic obtained from bacterial cultures, used for hair spray containers. These substances are several times more costly than the plastics currently used for packaging.

7.2.3.2. Chemical Recycling

In chemical feedstock recycling, the polymer chain molecules are broken down into smaller hydrocarbon molecules, which are fed to polymerization reactions. The process leads from polymers to monomers or, finally, back to oil.

With regard to the energy balance (Fig. 38) as well as costs, these processes fall between pure remelting and combustion.

There are two methods by which polymer molecules are cracked:

1) Chemical degradation by hydrolysis or hydrogenation
2) Thermal degradation (Fig. 39), where the molecules are disintegrated by heat alone (as in pyrolysis) or with addition of oxygen, which lowers the process temperature.

Catalysts are necessary or at least advantageous in each case.

As Figure 39 shows, this depolymerization or degradation leads directly to the monomer in some cases, and to a mixture of liquid or gaseous hydrocarbons in others. These low molecular mass substances can be readily separated and purified by standard methods and used as new feedstocks.

7.2.3.2.1. Chemical Degradation

Hydrolysis, Methanolysis, and Glycolysis. As early as 1981, hydrolysis was one of the topics investigated in the VKE research project [160], [161]. This process makes use of the fact that polycondensation, the reaction used in making polyamides and polyesters, is an equilibrium reaction in which at elevated temperatures the presence or absence of water governs whether the chains grow or are degraded. If water is added to a polycondensate at a high temperature, the compound decomposes to its monomers. A system for carrying this process, comprising a twin-screw extruder in which water or steam is injected into the polymer melt, is illustrated in Figure 40.

In the related process of methanolysis, a polymer such as poly(ethylene terephthalate), PET, is melted and mixed with methanol; depolymerization takes place between 160 and 240 °C. Feedstock recovery in the process is 99 %.

In the glycolysis process, PET is treated with an excess of glycol over catalysts and broken down into larger molecular fragments.

The alcoholysis of PET, polyurethane, and polyamide wastes from automaking has undergone a research renaissance [162]. An interesting potential application is the steam hydrolysis of foams from old auto seats. In the same way, the important bottle material PET can be made into new PET that is suitable for the same use.

This process can be performed on all polymers that contain carbonyl groups, although the monomers are not recovered in all cases. For example, diisocyanates are not obtained from polyurethanes.

Hydrogenation. When liquefied coal or liquid coal derivatives such as hydrocarbons are treated with hydrogen at 450 – 500 °C and 200 – 300 bar, the product is a mixture of a small amount of gas, much naphtha and oil, and some bitumen. This hydrogenation process (Fig. 41) was used to make high-quality fuels from coal in Germany during World War II. It yields a higher proportion of valuable liquid hydrocarbons (listed in Fig. 42) than does pyrolysis, although the plant is more expensive to build and operate. Of all the degradation processes, it is energetically the most favorable [163] because valuable oils and naphtha are generated preferentially. One of the few plants in the world using this liquid-phase hydrogenation process is the Kohle-Öl-Anlage facility at Bottrop, which takes in 25 t/h of heavy oil from the nearby

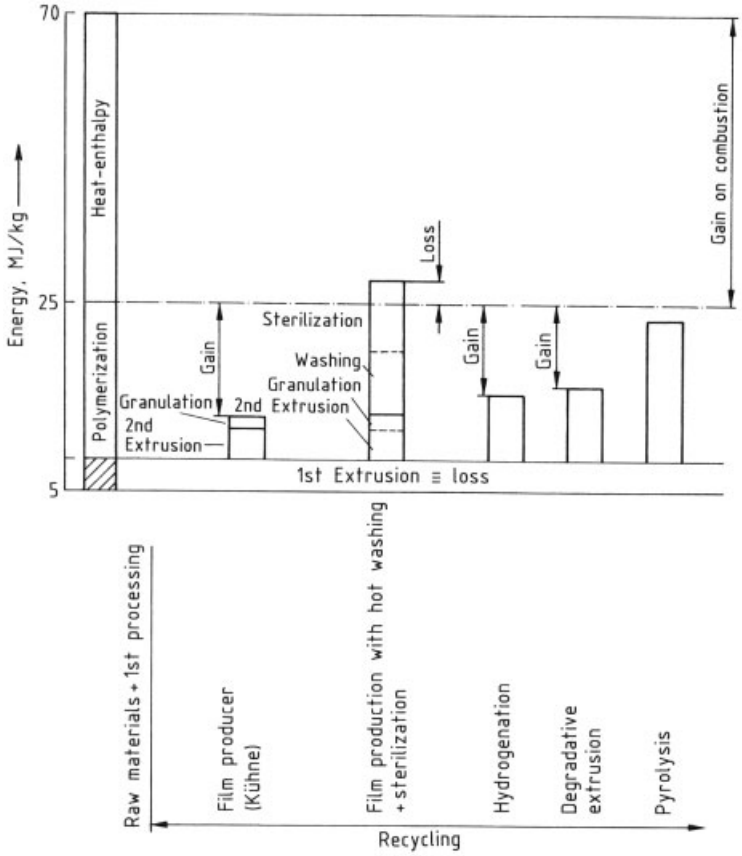

Figure 38. Energy balances for several processes involving polyolefins, with the enthalpy shown for comparison

VEBA refinery and produces naphtha and synthetic crude oil that are returned to the refinery as feedstocks.

In 1992 this plant processed 2000 t of mixed plastics packaging wastes. The wastes were ground to < 1 mm, mixed with the heavy oil (up to 7 %), and processed with it. A particularly high quality syncrude was obtained.

After expansion of the plant and improvement in the preparation of plastics wastes, capacity for these wastes will grow to 40 000 t/a after summer 1993, and subsequently to 200 000 t/a. VEBA is contemplating the erection of four more plants in Germany; these will have larger throughputs than the pilot plant now on stream, so that a capacity of over 10^6 t/a may be available by the year 2000.

7.2.3.2.2. Thermal Degradation

Pyrolysis. Bonds in polymer molecules are broken at elevated temperature, the chain strength depending on the nature of the covalent bonds:

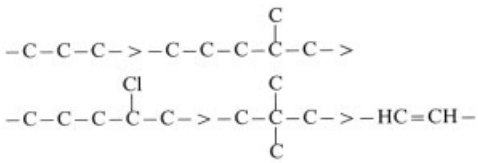

If air is excluded, chain cleavage begins at ca. 200 °C, and the molecular mass of the products decreases as the temperature rises: initially polymers, then hydrocarbons are obtained. This process is called pyrolysis [164].

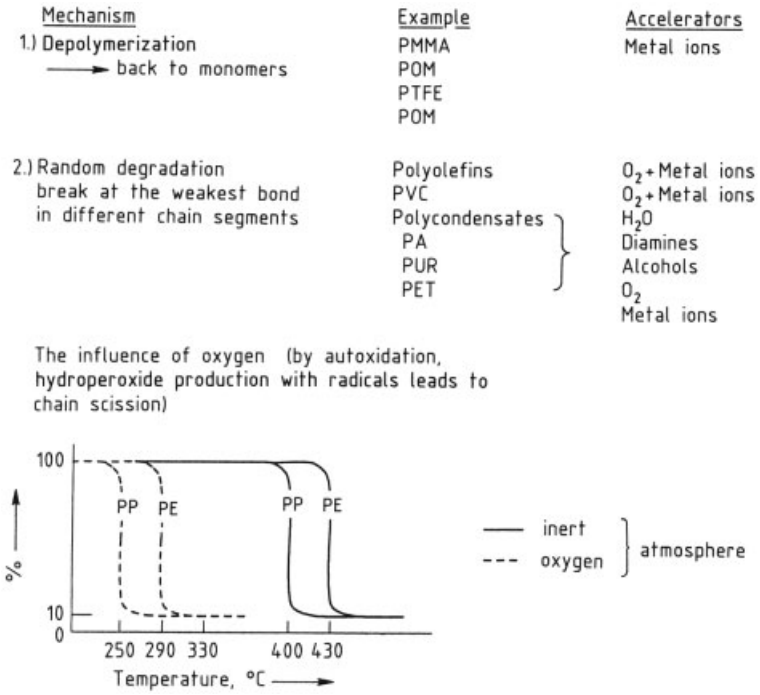

Figure 39. Options for depolymerization and degradation of resins

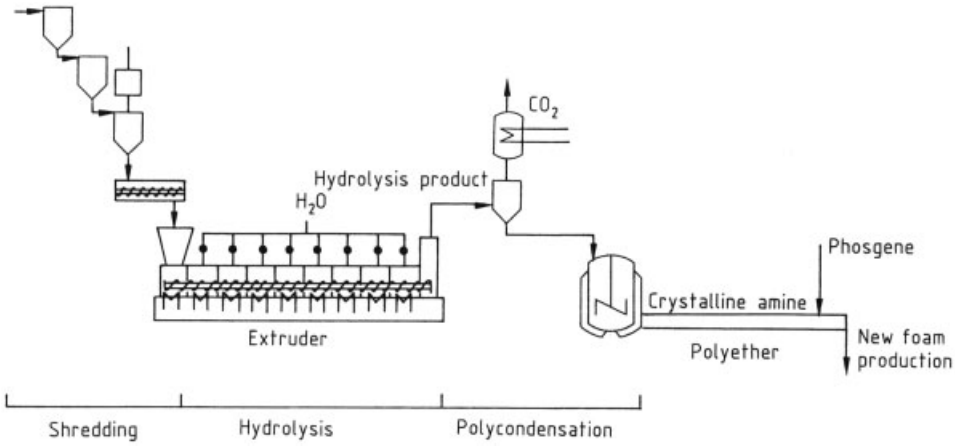

Figure 40. System for depolymerization by hydrolysis (Bayer) under the action of steam, shearing, and high temperature and pressure (250 – 300 °C, 25 bar, 5 – 20 min residence time)

Most of the plants now in operation are pilot plants [165] not specially designed for plastics but able to handle general refuse [166]. Because the goal is to obtain pyrolysis gas for use as heating fuel, these systems are operated above 1000 °C, so that the macromolecules are completely broken down.

If higher hydrocarbons and useful oils are to be made from plastics wastes, the plant must operate at ca. 700 °C (Fig. 43). PVC is separated from the feed, since chlorine poisons the product oils and would necessitate an expensive downstream dechlorination step. Unfortunately, pyrolysis yields large amounts of gaseous hydrocarbons, which can only be used for process

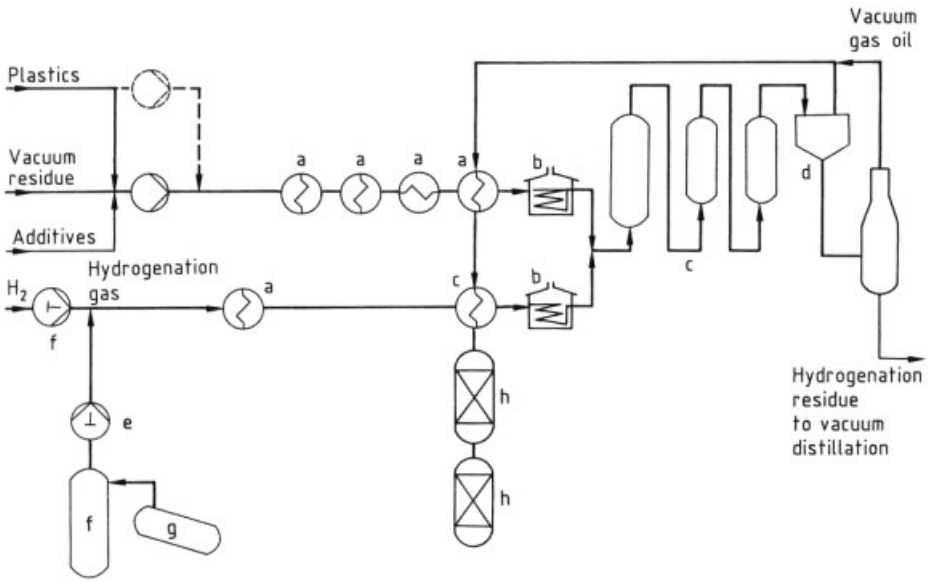

Figure 41. Simplified flowsheet of a hydrogenation plant (KAB)
a) Heat exchangers; b) Preheater; c) Bottoms-phase reactors; d) Hot flash drum; e) Compressors; f) Oil scrubber; g) Separator; h) Gas-phase reactors

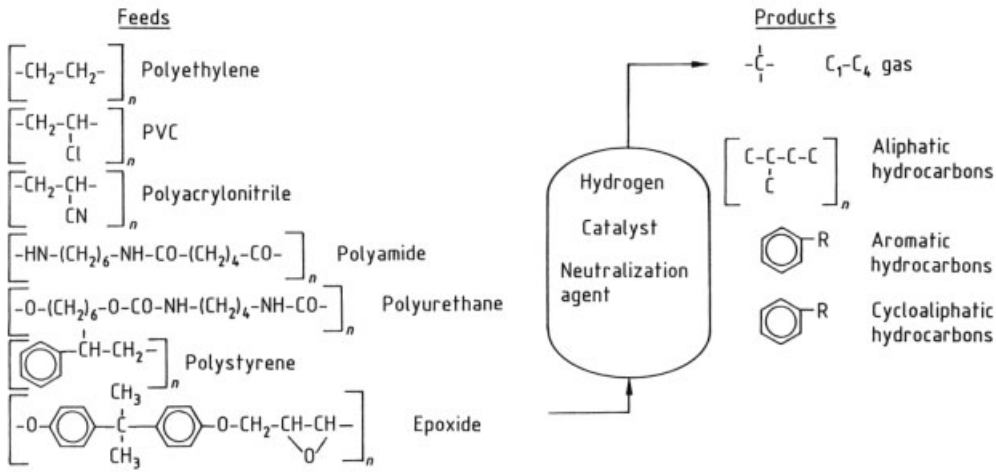

Figure 42. Products obtained by liquid-phase hydrogenation (KAB)

heating unless the plant is part of an integrated manufacturing complex.

This process is thought to be promising for elastomer wastes (such as auto tires) because they can be inlet to the process without grinding.

Degradative Extrusion. In degradative extrusion [167] a mixture of diverse polymers is treated at high temperature and under shearing, with reactive gases such as oxygen (air) and hydrogen (water), and in some cases catalytic additives such as metal oxides. The plastics are rapidly degraded to a homogeneous melt of low molecular mass polymers. Figure 44 shows two plants based on standard extruders. The first stage in the process removes chlorine, which originates chiefly from PVC in mixed plastics wastes. By suitable screw design, this is achieved

solely by heat and shear for PVC contents of up to 100 %; the chlorine content of the end product is < 0.3 %.

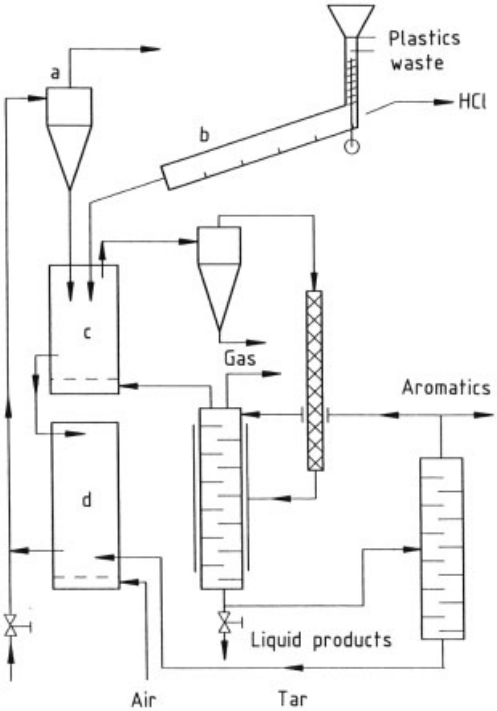

Figure 43. Fluidized-bed pyrolysis plant (after KAMINSKY) a) HCl absorption; b) HCl separation; c) Pyrolysis; d) Incineration

The mixture can also contain up to 40 % elastomers and thermosets, which are pulverized in specially designed screws. At ca. 400 °C, the discharge from the extruder is an oily liquid, which at room temperature solidifies to a brittle wax, suitable for various uses (Fig. 45). The oil, for example, can be fed directly to a hydrogenation reactor, a synthesis gas unit, or a blast furnace (forming a reaction gas for metallurgical processes; see Fig. 46). The extruder product can also be made into a conveyable powder or pellets. It would thus be possible to convert the wastes into an appropriate homogeneous form at the point of waste generation using a single extruder. The powder, pelletized material, or melt can then be forwarded to the actual processing plant (e.g., pyrolysis or hydrogenation) or to a refinery where it is mixed with crude oil or heavy oil fractions for processing in a cracker or hydrogenator. The material can also be stored indefinitely without danger. If the heavy oil is to be used as a refinery feedstock, the chlorine content must be < 20 ppm; work is now under way to attain this level.

A degradative extrusion plant for 1 t/h of plastics wastes costs ca. € 1.4×10^6. The cost of operating the process, including chlorine separation but not chlorine disposal is estimated at ca. 100 €/t. With equipment now available, plants with capacities of up to 20 t/h are conceivable.

Partial Oxidation. In the partial oxidation process, heavy oil is converted to synthesis gas, a feedstock for the production of methanol, urea, and other basic chemicals. The heavy oil is heated to 250 °C and mixed with oxygen and steam, then atomized into a reactor. Part of the oil burns, supplying the heat to bring the process up to 1300 – 1500 °C. The product gas mixture consists of ca. 45 % carbon monoxide and ca. 45 % hydrogen, with 10 % carbon dioxide, which must be removed by scrubbing.

It would also be possible to replace the heavy oil with a synthetic heavy oil made from mixed plastics wastes in an degradative extruder, fluidized-bed reactor, or melting tank. Figure 47 shows schematically how a Shell process syngas plant might be combined with a degradative extrusion system handling mixed plastics wastes.

Process engineers at Farbwerke Hoechst have developed a fluidized-bed partial oxidation process that operates at 750 – 800 °C. The products are designed to be basic chemicals needed in large quantities in the chemical industry (Fig. 48). The results of the calculation show that this process is only partially competitive with landfills. The ESPAG plant at Schwarze Pumpe, using reactors originally designed for briqueted lignite, has already processed 20 000 t of waste plastics (packaging) together with lignite briquetts to produce town gas. The company is now converting its plants to make methanol from the gas [168].

Total Oxidation. Organic substances, such as mixed organic wastes that also contain plastics (e.g., household waste or shredder waste), can be treated with oxygen at high temperatures (>1500 °C) and completely converted to carbon dioxide and water vapor without any residual toxic hydrocarbons remaining. Figure 49 shows a closed cycle such as might be achieved with

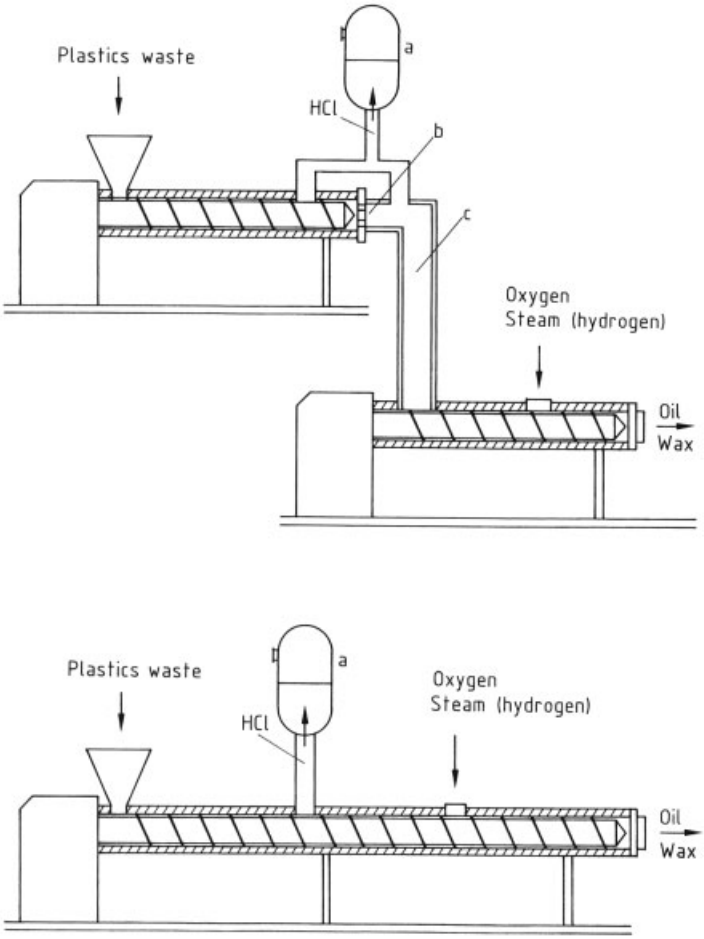

Figure 44. Degradative extrusion
a) Neutralization scrubber; b) Filter; c) Vacuum

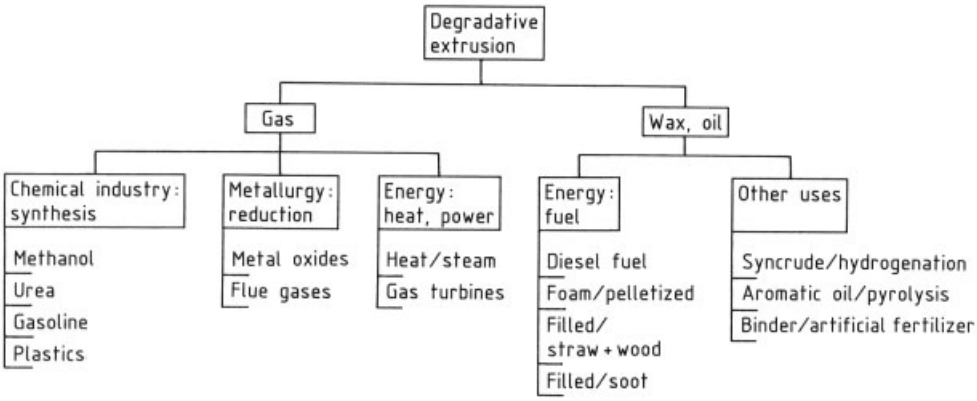

Figure 45. Range of products available from degradative extrusion of mixed plastics wastes

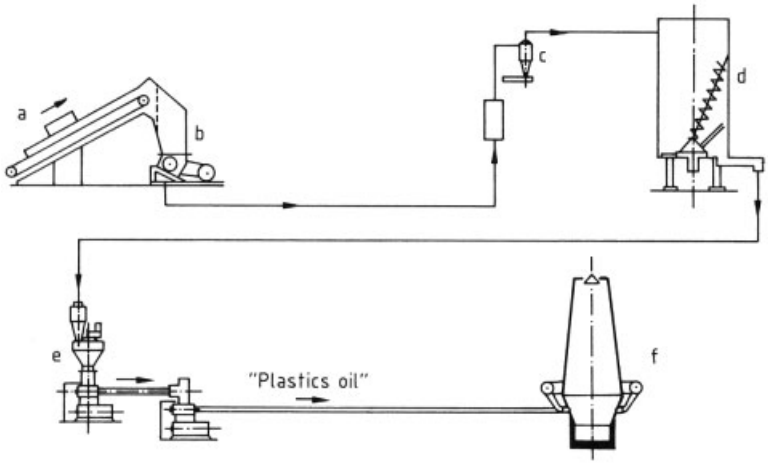

Figure 46. Degradative extrusion to produce plastics-derived oil for blast-furnace reduction or iron ore (Niemöller, Klöckner)
a) Charging; b) Grinding; c) Separator; d) Bunker; e) Extruder; f) Blast furnace

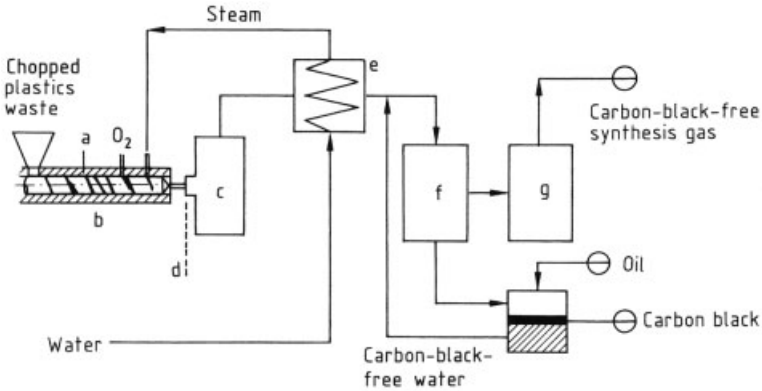

Figure 47. Combination of Shell gasifier and extruder
a) Sensor; b) Barrel (400 °C); c) Reactor; d) Burner; e) Waste-heat boiler; f) Carbon black scrubber; g) Gas scrubber

such a process. By operating under substoichiometric conditions and metering hydrogen or hydrogen-rich substances into the off-gases, a synthesis gas can be made by reduction; this can be used in producing basic chemicals such as methanol. Some possible hydrogen-rich compounds (besides used oils) are polyolefin wastes liquefied by degradative extrusion.

Figure 50 shows a schematic of how such a plant might be built [169]. This system is especially suitable for converting mixed and contaminated plastics- containing wastes without pretreatment. It is equipped with a rotary kiln that is insensitive to melting thermoplastics. Because the combustion gas is pure oxygen, the volume of off-gas is only ca. 25 % of that in an air combustion plant, chiefly because nitrogen is not present. The combination has the following advantages:

1) The volume of pollutants is reduced to ca. 25 % compared to air as combustion gas
2) The off-gases have a high concentration of carbon dioxide (nearly 50 %), which permits the gas to be recovered and converted to synthesis gas, as shown in Figure 49.

The rotary furnace in Figure 50 functions as the oxidation reactor and is linked to a chemical plant via the gas- cleaning stage (removal of sulfur and chlorine compounds). In the example shown, the downstream plant makes urea solution.

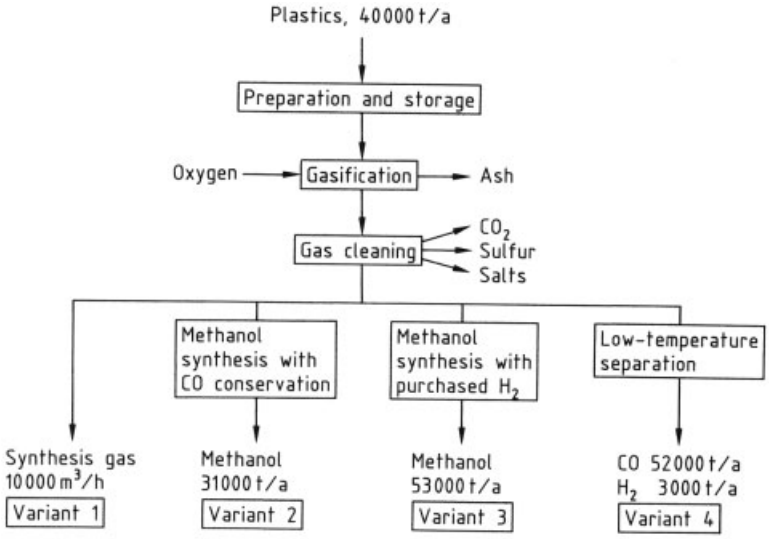

Figure 48. Fluid-bed gasification (Hoechst)

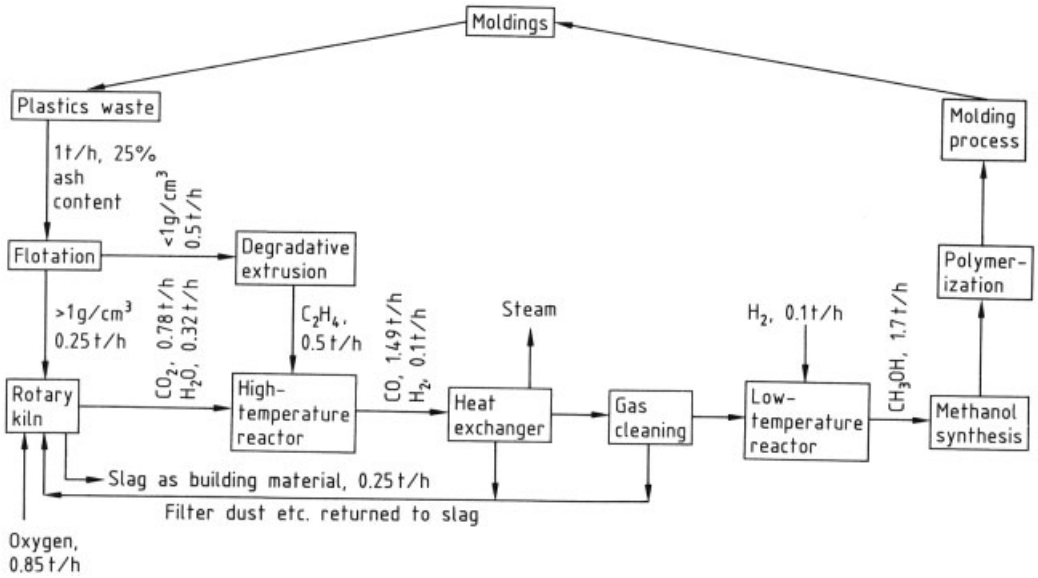

Figure 49. Total oxidation of plastics wastes with reduction of tail gases in order to produce synthesis gas for synthesis of methanol and polymerization of polyoxymethylene (closed cycle)

7.2.4. Combustion (Energy or Thermal Recycling)

The expression "energy recycling" refers to utilizing the heat content of plastics wastes. Common plastics are an excellent fuel; polyolefins have a higher heating value than fuel oil (see Fig. 51), are cleaner, and contain no toxic or unpleasant species such as sulfur. Because polyolefins account for ca. 65 % of plastics in household waste and the other plastics in such waste also have high heating values, the average 7 % content of combustible plastics enables waste incinerators to save ca. 20 % of the fuel oil that would otherwise be needed for auxiliary firing. The proposition that it is economically desirable to utilize at least the energy of the plastics wastes contained in refuse is not disputed among

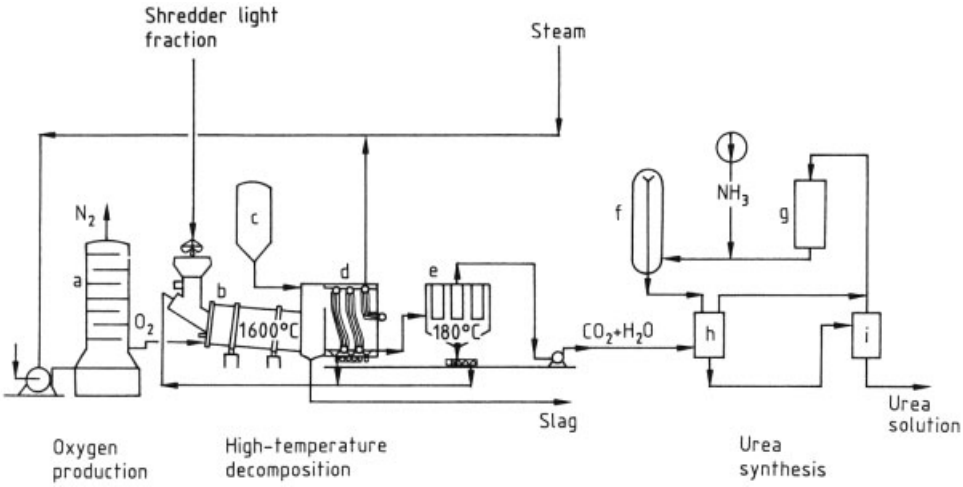

Figure 50. Combination of combustion with pure oxygen and chemical plant utilizing synthesis gases, for the conversion of shredder light fraction to urea solution
a) Air distillation unit; b) Rotary kiln; c) Additive bunker; d) Waste-heat boiler; e) Fabric filter; f) Reactor; g) Condenser; h) Stripper; i) Decomposer

experts. This is even regarded as an alternative energy source for the future [170], [171]. Furthermore, this form of waste disposal cuts the volume shipped to landfills to ca. 10 % when conventional combustion is used. By using high-temperature combustion plant with oxygen and, e.g., a rotary kiln, a vitrified slag can be obtained that even retains heavy-metal oxides in noneluable form. Such slags can be used as construction materials, reducing the volume deposited in landfills to < 1 %.

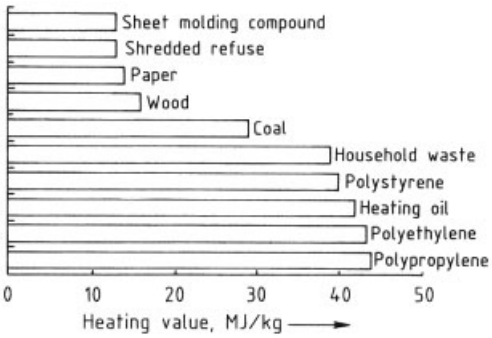

Figure 51. Comparative table of heating values

7.2.4.1. Combustion in Waste Incinerators

The 7 % content of plastics in household waste is viewed by noted experts [172] as desirable for the operation of modern waste combustors, since the high heating value makes it possible to economize on other energy sources. Some 20 % of the fuel oil required as an auxiliary heat source for unsorted household waste can be saved [172], [173].

7.2.4.2. Shredded Waste as Fuel for Sewage Sludge

Tests in Bayer's rotary-kiln waste incinerator [174] have shown that the organic components of the shredder light fraction have a high heating value of > 18 MJ/kg, 50 % of that of coal. Burning this fuel produces no dioxins or other detrimental effects. Nevertheless, because of contamination with toxic used oil, shredder waste from automobiles is barred from normal waste-combustion facilities in Germany. Therefore, the concept being pursued is to mix shredded waste with sewage sludge and burn them together [175]. Tests have shown that this can save the fuel oil otherwise required for sludge combustion [176]. Heavy metals are fixed in the ash together with other toxic components already present there, albeit in a readily recoverable form. After the valuable heavy metals have been reclaimed, however, the ash residues must be dumped in landfills.

7.2.4.3. Combustion with Pure Oxygen [177]

Rotary kilns have proved best for the combustion of thermoplastics or wastes that contain large amounts of thermoplastics. If the furnace is fed with oxygen instead of air, the plant, and especially the flue-gas cleaning unit, can be downsized due to the absence of nitrogen. The reduction in volume by $> 75\,\%$ has other advantages, such as a very high adiabatic temperature, better efficiency, and less residue for landfill. Because of the advanced flue-gas cleaning techniques now used, combustion with oxygen ensures that emission standards will easily be met. A plant based on this concept corresponds to the combustion reactor on the left-hand side of Figure 50. Figure 52 shows an energy and mass balance for a plant of this kind processing 1 t/h of shredder light waste (18 MJ/kg). The heat generated in the system is utilized to heat the liquid oxygen for the cooling and condensation of CO_2, which is produced for industrial use.

Because the high adiabatic temperatures involved in combustion with concentrated oxygen necessitate cooling of the furnace, which can be accomplished by admitting unheated wastes or by recycling the stack gas, sludge from gas cleaning and dust from the filters are returned to the furnace and thereby incorporated in the slag, thus further reducing the volume going to landfills.

Other pollutants, such as gaseous chlorine and sulfur compounds must be removed from the flue gas before it is released to the atmosphere.

The energy required to produce oxygen is $0.5\,kW/m^3$, which is roughly half of the electrical energy generated in 1 t/h plants. Liquid oxygen costs $0.09\,€/m^3$ O_2 for major consumers. If the energy to make liquid oxygen is available, the price drops to $0.02\,€/m^3$.

Extra costs (disposal costs) for the combustion of shredded waste in plants of this kind are estimated at 130 and 230 €/t.

The process would offer the following advantages:

1) Due to the very high temperatures, a glass-like liquid slag is produced. Toxic heavy-metal oxides are fixed in noneluable form in this slag, which can therefore be used as a building material. Glass wastes would thus be desirable in the wastes fed to the process.
2) The volume of stack gas produced is only about a third as much as in combustion with air. This means that
   a) Investment costs are lower
   b) Very high temperatures are attained without supplementary heating
   c) Operating costs are low
   d) There are no emissions that would pollute atmosphere, soil, or water

7.3. Solvents

7.3.1. Introduction

Origin, Quantities, and Composition of Solvent Wastes. Organic solvents are employed in many industrial processes: in extraction and dispersion, extractive distillation, gas chromatography, chemical reactions, and recrystallization; in the production of paints, pickling agents, printing inks, rubber solutions, films, synthetic fibers, and plastics and resin solutions; for degreasing and chemical textile cleaning; and as aerosol propellants.

According to a nationwide examination of shipping documents for hazardous wastes, the quantity of organic solvents and solvent mixtures used in Germany in 1987 was ca. 385 000 t. Among the substances not included in this figure are cold cleaning agents and plasticizers. Some 38 % of the total represents halogen- containing solvents or solvent mixtures, while ca. 62 % is halogen-free [178].

Because of the wide variety of solvents and solvent applications, solvent wastes vary widely in composition as well as nature and content of impurities. Furthermore, used solvents are often collected in mixed form.

Environmental Relevance of Solvent Wastes. A large proportion of organic solvents are "hazardous wastes" as defined in § 2 Section 2 of the Federal Waste Act (Abfallgesetz) of August 27, 1986 [179] and are thus subject to accountability. Such substances pose hazards to human health, the atmosphere, water, and soil and are explosive or combustible.

Humans absorb organic solvents mainly via the respiratory tract and the skin. The resulting

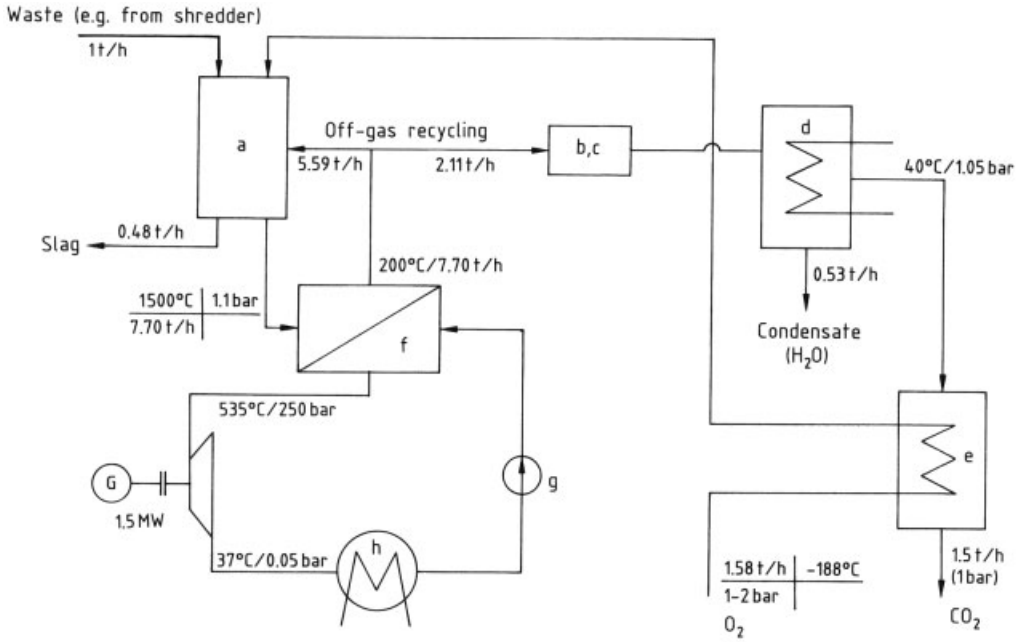

Figure 52. Energy and mass balance of a shredder waste (18 MJ/kg) incineration plant operating with pure oxygen (Renz, Bu, Doll)
a) Combustion; b) Sulfur removal; c) Nitrogen removal; d) Aftercooler; e) O_2 preheater; f) Steam generator; g) Pump; h) Condenser

effects depend on the solvent itself, the quantity present, and the time for which it acts. Short, high exposures can lead to acute harm to the organism. The intake of small quantities over long spans of time can cause chronic diseases.

Solvents are absorbed and accumulate in fatty cells (nerves, brain, bone marrow, body fat, liver, kidneys), causing damage specific to the solvent and organ.

7.3.2. Processing of Solvent Wastes

7.3.2.1. General Aspects

Legislation requires that operators of facilities subject to regulatory approval minimize residues through abatement and recycling practices (§ 5 Section 1 No. 3 of the Federal Pollution Control Act, BImSchG [180]. Abatement practices include process optimization and/or modification as well as internal treatment and recycling units. External recycling takes place in central facilities. The recyclates are forwarded to the producer or returned to the economy as market chemicals.

Unmixed solvents with low impurity contents can generally be treated more readily than solvent mixtures with high concentrations of impurities. The prerequisite for effective, economical recycling of solvents is therefore that used solvents be collected separately. They must at least be separated into halogen- containing and halogen-free solvents, but it is best if solvents are collected individually.

A regulation concerning the acceptance and recycling of used halogenated solvents by the seller came into force on January 1, 1990 [181]. This regulation mandates that users of solvents collect residues separately and bans the mixing of halogenated with halogen-free solvents.

Methods used to purify waste solvents and solvent mixtures include:

Mechanical processes
Screening
Filtration
Sedimentation

Thermal processes
Distillation

Azeotropic distillation
Vacuum/pressure distillation
Hydrogenation
Gasification
Thin-film evaporation
Condensation
Chlorolysis
Extraction
Adsorption/absorption

Physicochemical processes
Freezing-out of water
Neutralization
Desiccation

The processes are essentially the same as those used to refine crude oil by removing solids, drying, eliminating inorganic components (salts), and separating into organic components or substance classes. Detailed process descriptions of methods can be found in [182–184]. Methods specially employed for the recovery, treatment, and recycling of solvents are summarized in [178], [185], [186], and elsewhere.

7.3.2.2. Mechanical Separation of Solids

Techniques used to separate solids and liquids include filtration, sedimentation, and centrifugation.

Filter presses and multicompartment drum filters are proven devices for separating solids from organic liquid systems. Simple gravel filters and centrifuges are less widely used. Equipment is fully described in the literature [187], [188].

In filter presses the slurry to be separated is inlet to individual chambers formed by frames and perforated plates that can be moved relative to one another. The plates are covered with filter cloths. The liquid phase can escape only by passing through the filter cloths to the discharge. The solids remain as a cake on the cloths. Separation can be speeded up by squeezing the filter plates together. When the filter cake has reached a certain thickness and interferes with or prevents further filtration, the press must be opened and cleaned. Thus filter presses can only be operated in batch mode.

Drum filters (Fig. 53) are more suitable for continuous operation. The drum, divided into chambers and covered with a fabric filter, rotates in the filter tank containing the slurry. A special valve imposes a vacuum on the chambers that are in the slurry. The liquid phase passes into the chambers and is drained off, while the solids remain on the filter medium to form a cake. When the filter chambers rotate out of the tank, the cake can be washed and dried by vacuum. It is then removed by a scraper. Compressed air or steam can be applied to the chambers to aid separation. After cake removal, the fabric may be washed. The cycle then begins over again.

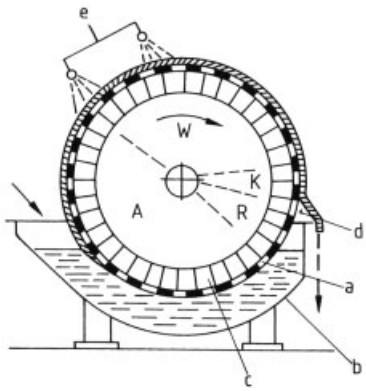

Figure 53. Design and operation of a drum filter [188, Vol. III]
a) Filter drum; b) Filter tank; c) Compartments; d) Cake discharge; e) Washing sprays
Zones: A) Vacuum, W) washing and dewatering, K) cake discharge, R) cleaning

7.3.2.3. Treatment of Halogenated Solvents

Highly chlorinated solvent wastes occur chiefly in the chemical industry. The following processes for their recovery and disposal are now installed or under development [189]:

1) Chlorination or chlorolysis to chlorinated hydrocarbon solvents
2) Solvent distillation
3) Dechlorination to hydrocarbons
4) Hydrogenation
5) Combustion
6) Physicochemical treatment

Of these processes, chlorolysis is most important economically. Chlorolysis converts the unavoidable residues of chlorinated hydrocarbons into valuable substances. Products with a low

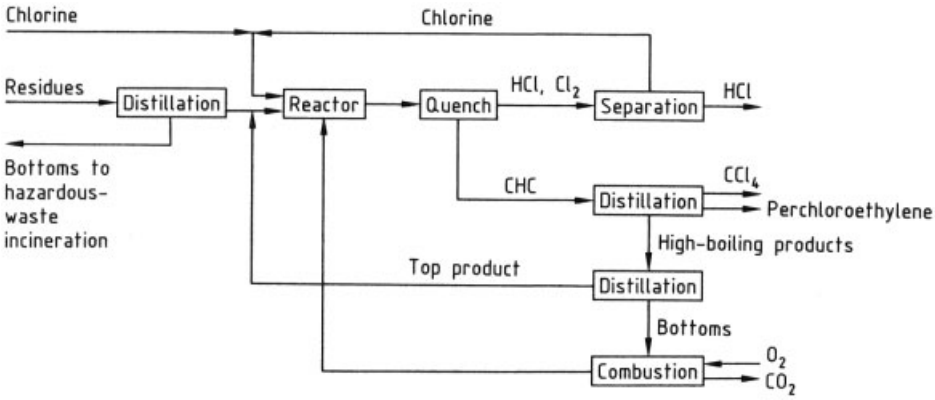

Figure 54. Flowsheet of the Hüls low-pressure chlorolysis process [190]
CHC = chlorinated hydrocarbons.

degree of chlorination, for example from pharmaceutical and fine chemicals manufacturing, generally cannot be recycled by this method.

Numerous process variants have been developed for low-pressure chlorolysis producing tetrachloromethane and perchloroethylene simultaneously. Figure 54 shows a schematic of the Hüls low-pressure chlorolysis process.

Chlorinating cracking is performed in tubular or fluidized-bed reactors. The residues are high-boiling products, from which hexachloroethane and hexachlorobutadiene can be separated and recycled, and hydrogen chloride.

Chlorofluorocarbons (CFCs), used as refrigerants in air conditioning and cooling systems, represent a serious threat to the atmosphere because they damage the ozone layer. These substances can be recycled only if they are unmixed. Mixtures of refrigerants cannot be regenerated and must be disposed of in a hazardous-waste incinerator.

Before they are separated by distillation, refrigerants must be dried and deacidified. Figure 55 shows a schematic of a refrigerant purification process.

The high-boiling impurities are removed first. Acids are next eliminated and the product is dried before the gas phase is condensed and the condensate inlet to distillation. Byproducts and residues from the distillation step are inert gases and high-boiling constituents; these are forwarded to combustion plants together with other residues from the process. The regenerated refrigerant is led to a storage tank via a filter.

CFCs may not be produced, used, recycled, and placed back on the market after January 1, 1995. Global, progressive environmental thinking therefore demands that the focus should be shifted to safe destruction of these substances by incineration.

7.3.2.4. Treatment of Halogen-Free Solvents

Halogen-free solvents are becoming increasingly important as substitutes for halogenated solvents. The recovery and recycling of solvents from the gas phase depend on many factors (Fig. 56). The figure and the discussion below relate to halogen-free solvents and should apply by analogy to those containing halogenated products.

Processes for recovering solvents from the gas phase are based on the following principles:

1) Condensation
2) Adsorption on solids (activated carbon, silica gel, alumina, molecular sieves, polymers)
3) Absorption in liquids
4) Separation according to molecular size (membrane processes)

Adsorption on activated carbon and condensation have become particularly important. After adsorption the solvent is recovered by desorption, or the solvent-laden carbon is combusted. Combustion is generally practiced when gases with low solvent contents are cleaned and thus regeneration is not economical, or when many different solvents are collected together.

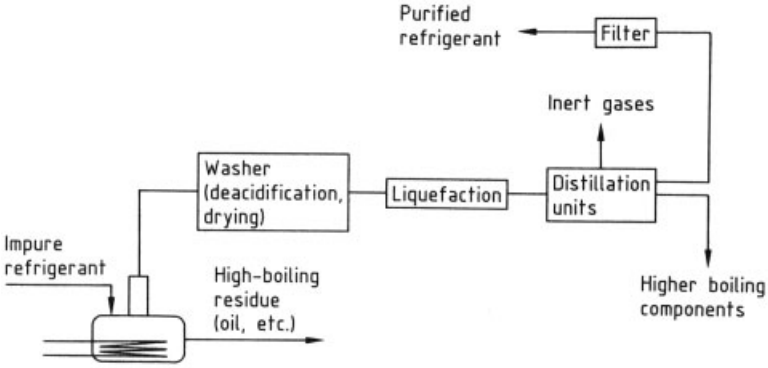

Figure 55. Flowsheet for purification of refrigerants by the Hoechst process [191]

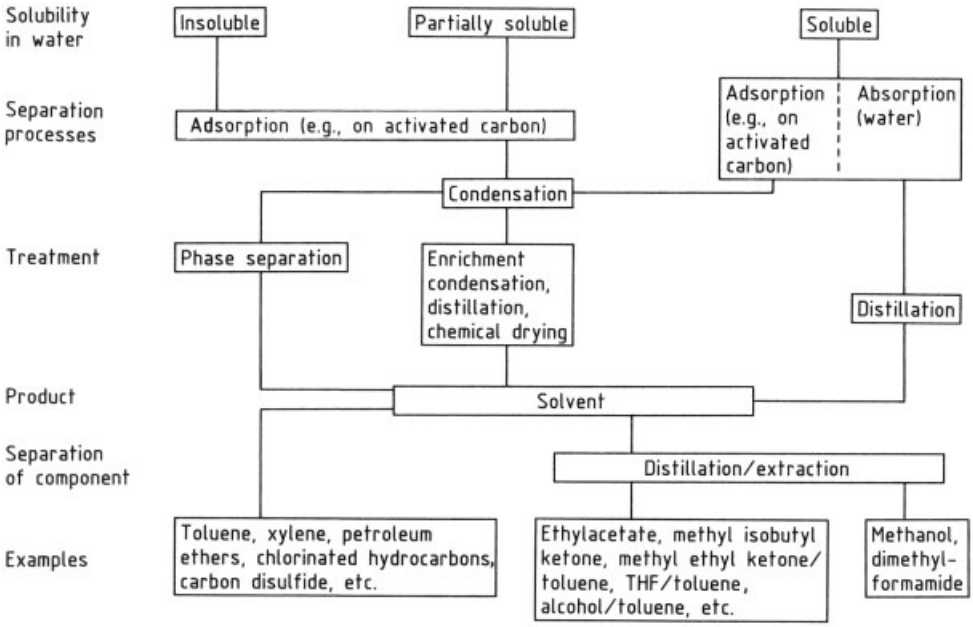

Figure 56. Influencing factors and possibilities for recovery of organic solvents [192]

The methods used to separate adsorbent and solvent include steam desorption, vacuum desorption, and inert-gas desorption.

When steam is used for desorption, the process must include a water/solvent separation stage. Some methods are fractional distillation at atmospheric pressure or under vacuum, vacuum distillation with direct vapor compression, rectification in a thin-film evaporator, and batch distillation.

In the area of waste-air cleaning and solvent recovery, there are many process developments optimized for particular purification and separation problems, and for particular upstream and downstream processes [193], [194].

Fischer has devised a method for cleaning waste air and recovering all types of solvents [195] in an absorption – desorption plant. The scrub liquor and absorbent is heat-transfer oil, which collects the solvents from the gas phase. The solvent-laden oil is regenerated in the desorber, a specially designed vacuum distillation unit.

Halogen-free solvents and solvent mixtures arise chiefly from cleaning operations in print shops, painting and coating plants, and so

forth. These solvent wastes contain paint, resins, and other substances whose adhesive properties cause problems in treatment and may even render recycling impossible.

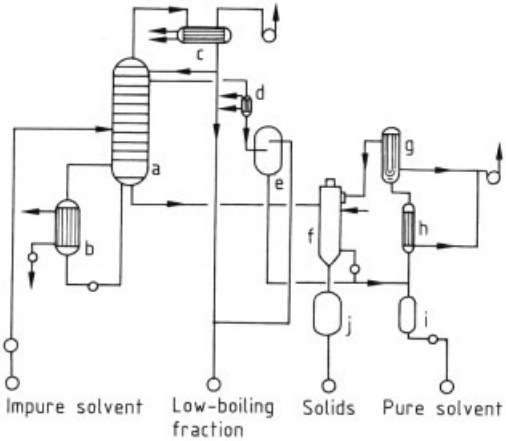

Figure 57. Distillation system for solvent recovery [197] a) Distillation tower; b) Forced-circulation evaporator; c) Condenser; d) Cooler; e) Separator; f) Drying tower; g) Condenser; h) Cooler; i) Solvent tank; j) Solids receiver

Contaminated solvents with low solids contents can be treated in a forced-circulation evaporator and concentrated to ca. 25 % solids. The circulation rate in the evaporator must be great enough to prevent blockage of the tubes.

For high solids contents, the standard distillation method may be unsuitable. Adhesive contaminants may have to be removed in thin-film evaporators or stills where vaporization takes place through a semipermeable layer that does not allow the contaminants to pass.

A thin-film evaporator consists of a heated cylindrical shell with an internal rotor that distributes the product as a thin film on the heating surface and scrapes the evaporation residues off the heating surface. The solvents evaporate and are collected in a downstream condenser. Residues are drained off at the bottom of the evaporator.

The processes used for solvent wastes free of adhesive contaminants include:

1) Mechanical recovery
2) Evaporation
3) Distillation (including single- and multistage, steam, vacuum, and azeotropic distillation)

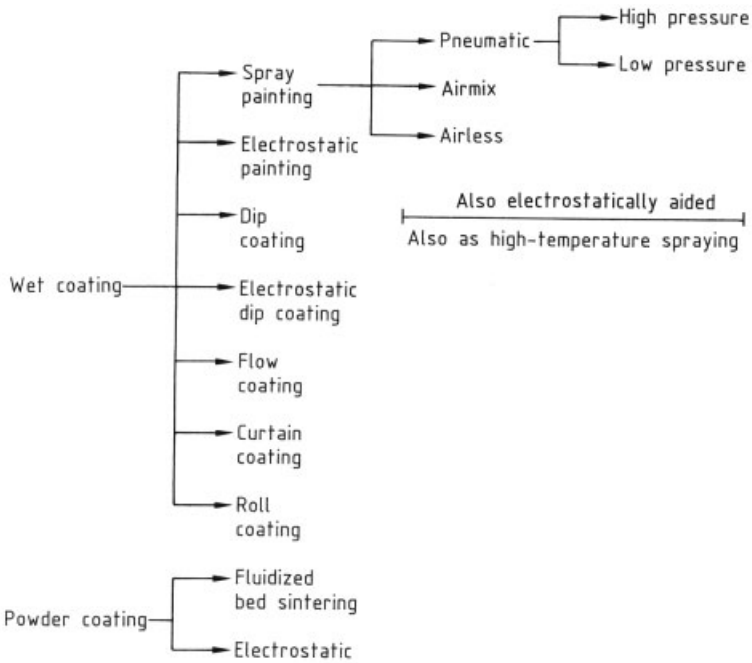

Figure 58. Summary of painting processes

Often, a combination of these methods is necessary [196]. Figure 57 shows a distillation system.

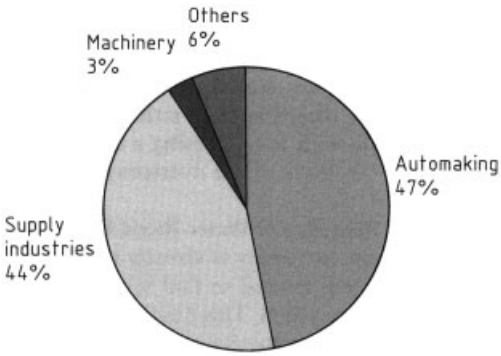

Figure 59. Paint sludge by industry of origin in the Land Hesse, Germany, 1989

7.4. Paints

7.4.1. Introduction

Origin, Quantities, and Composition of Paint Sludges and Paint Wastes. Articles made of steel, plastics, wood, and other materials are coated with paints and varnishes in order to preserve them. The residues from coatings and coating operations represent a potential environmental hazard. These residues include paint sludges and paint wastes.

Paint sludges are produced whenever exhaust air laden with overspray is cleaned up by wet scrubbing (coagulation). More than two-thirds of the residues from painting installations are paint sludges [198].

Paint wastes are uncoagulated, dried, or dissolved paints originating at drainage (drip) points, as surplus, through improper or prolonged storage, or in cleaning operations.

The nature, quantity, and composition of these residues depend on the painting process and the branch of industry using the paint. Figure 58 lists the common painting techniques. Spraying with wet paints remains the most widely used technique. Comparatively large amounts of paint sludge arise from spray painting installations, especially in the case of simply compressed-air spraying [199].

Figure 59 presents a breakdown of paint sludge production by industry for the German Land of Hesse in 1989. The amount of paint sludge from plants requiring regulatory approval was 1625 t/a [200].

The 1987 tabulation of shipping documents for hazardous wastes in the Federal Republic of Germany gave a total output of ca. 144 000 t of paint sludge [201]. In Berlin, 3142 t of paint sludge was generated in 1991 [199].

Paint sludges vary widely in complexity and heterogeneity. They contain 50 – 70 % water and 1 – 10 % coagulant. The remainder is paint, which in turn contains vehicles, pigments, fillers, modifiers, additives, and solvents (see Fig. 60).

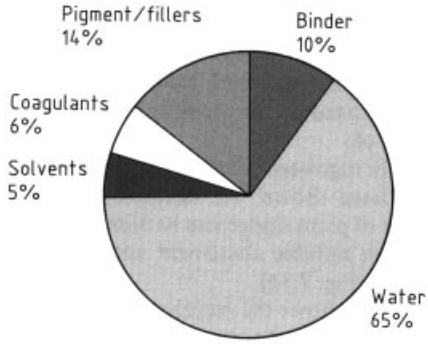

Figure 60. Composition of paint coagulate

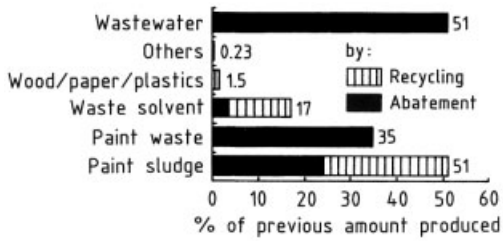

Figure 61. Potential for waste abatement from regulated painting facilities in Hesse [200]

Environmental Relevance of Paint Sludge. Paint sludge generally contains solvents and/or heavy metals, and is therefore hazardous waste. These residues are for the most part disposed of in hazardous-waste combustion plants or deposited in hazardous-waste landfills.

When paint sludge is incinerated, the heavy metals liberated from certain pigments necessitate special methods of off-gas and residue treatment.

Virtually all the solvents in paints are released as atmospheric pollutants. In the Federal Republic in 1986, a total of 380 000 t of organic solvents was released into the air. Next to automobiles and trucks, the use of solvents is thus the largest emission source of volatile hydrocarbons [202].

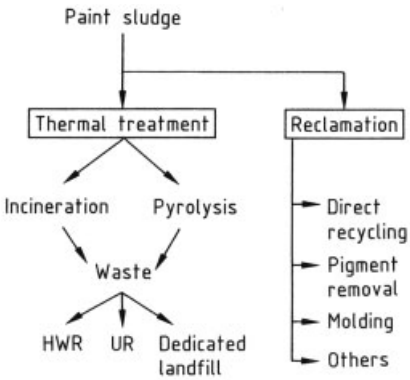

Figure 62. Recycling and processing of paint sludge
HWR: hazardous waste repository; UR: underground repository

7.4.2. Recycling of Paint Residues

7.4.2.1. General Aspects

Operators of plants subject to regulatory approval must cut down the amount of residues generated and/or recycle their residues (§ 5 Section 1 No. 3 of the Federal Pollution Control Act, BImSchG [203]).

The production of paints involves major expenditure on plant, energy, and oil-based raw materials. Even if mixed with water, paint sludges are therefore of high economic value. The loss of paint coagulate in the Federal Republic of Germany in 1990 was estimated at 200 000 t with a value of around € 1.2×10^9 [204]. To conserve this value, internal processes that allow direct reuse should be chosen whenever possible, or external recycling should be practiced.

In painting shops paint residues can be greatly reduced by replacing conventional spraying by more paint-efficient application techniques. If the processes cannot be changed, process optimization and internal and external recycling must be used to reduce the quantity of residues to be disposed of.

Studies in regulated painting plants in Hesse, Germany, have shown, for example, that the 1989 output of paint sludge can be diminished by 51 % through suitable abatement and recycling practices (see Fig. 61).

Figure 62 shows the recycling and disposal options. Deposition of paint sludges in landfills is no longer permitted [205].

7.4.2.2. Internal Paint Recycling

Direct Recovery with Collection Equipment. Overspray is intercepted by appropriate devices and removed from them by doctor blades, solvents are added, and the product is returned to the painting process. Belts and disks are proven types of collection equipment. These direct paint recovery methods are particularly effective in spray processes where the particles are subjected to high acceleration and their trajectories are highly stable in direction (e.g., airless spraying). Figure 63 shows a process using a disk, which can recover 70 – 90 % of the overspray.

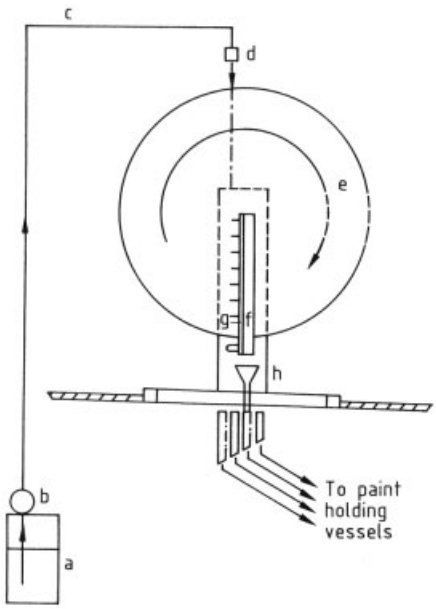

Figure 63. Disk-type overspray recovery system [198]
a) Solvent vessel; b) High-pressure solvent pump; c) High-pressure solvent feed line; d) Intermittent solenoid-actuated valve; e) Stainless steel disk; f) Stainless steel doctor blade; g) Nozzle assembly with cover; h) Replaceable funnel

Internal Recycling of Water-Based Paint. Water-based paint overspray is already intercepted by scrubbing and treated so that it can be recycled to the spray system. This kind of recycling is already in use for water-based baking enamels (which do not harden spontaneously) and has been implemented on a pilot-plant scale for airdrying water-based paints. Ultrafilterable latex paints for wood, commercially available since September 1992, permit internal recycling of water-based paints in the furniture industry [199].

The prerequisites for internal recycling are that:

1) The recirculating water contains no contaminants
2) The paint is "stable" (does not break down when highly diluted)
3) The fraction of overspray is not so great as to make the paint unusable (even small shifts in pH can cause coagulation)

The following processes (paint concentrators) can be used singly or in combination to separate the recirculating stream into a paint stream and a water stream:

1) Mechanical separation (decanters, centrifuges, etc.)
2) Ultrafiltration
3) Vacuum evaporation
4) Electrophoresis

Mechanical separators, however, are used only in combination with one of the other three methods.

The most important method for concentrating water-based paints is ultrafiltration. Vacuum evaporation and electrophoresis have serious drawbacks and, at present, can be used only in combination with other processes.

Figure 64 shows a flowsheet for internal recycling of water-based paint by ultrafiltration. The overspray is intercepted by a water curtain; no coagulant is used. The loaded scrub water is then fed to the filter unit, where it is separated into a low-paint substream, to be returned to the scrubbing process, and a paint-rich substream. The recovered paint is mixed with fresh paint and delivered to the automatic spray head. The maximum content of recycled paint is 35 – 45 %.

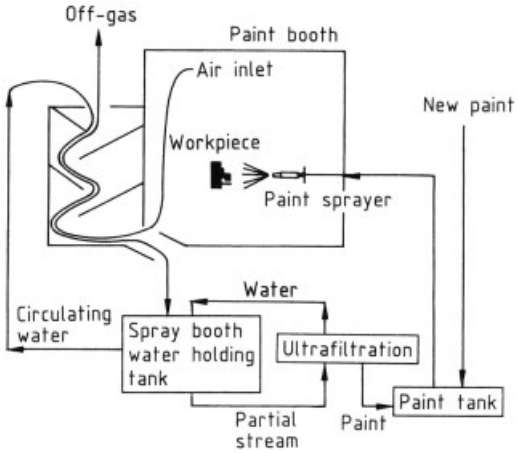

Figure 64. Internal recycling of water-based paint by ultrafiltration [199]

Figure 65 shows a more complex process for recycling of water-based paint in which a variety of concentrators are used. Undissolved paint particles are removed in the separator and returned to the paint tank (as recyclate 1). The liquid phase (thin paint) goes to the ultrafiltration unit, where it is enriched to a concentration suitable for electrophoresis.

7.4.2.3. External Paint Recycling

Recycling. The aim of recycling is to return the paint sludge to the process in one of the following forms:

1) Reconditioned paint
2) Base material for paint production
3) Base material for use in insulating materials, filler, moldings, and aqueous dispersions

The reconditioning of paint sludge to obtain reusable paint or base materials for paint preparation has the following prerequisites:

1) Neutral to weakly basic agents are used for coagulation to avoid denaturing the paint structure
2) Different paint systems are not mixed
3) Paint wastes are collected by color whenever possible
4) The scrub-liquor stream of the painting plant is kept free of other wastewaters and other wastes

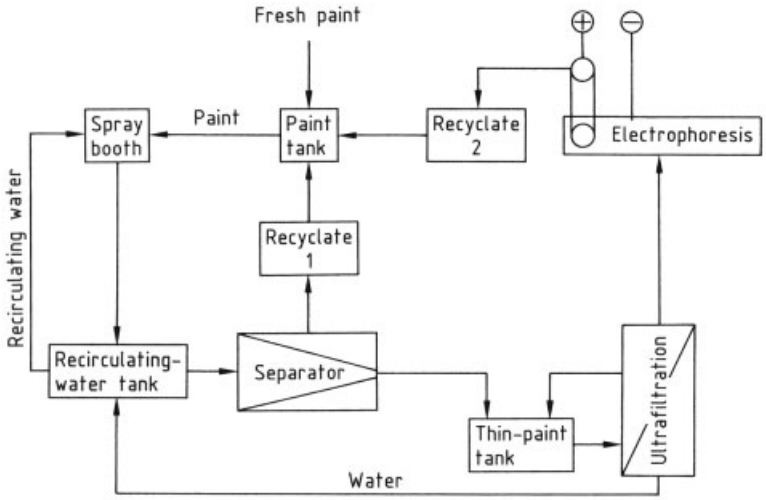

Figure 65. Complex water-based paint recycling system with various types of concentrators [206]

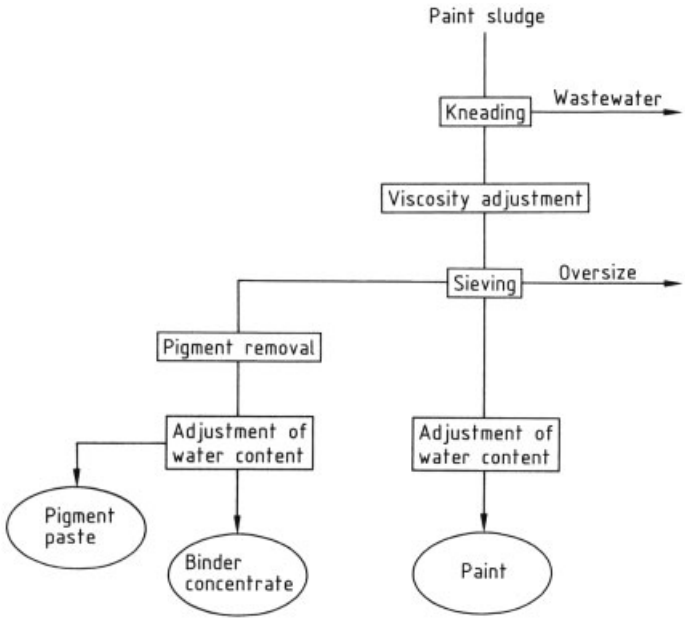

Figure 66. Flowsheet for reprocessing paint coagulates by the Envilack process [201]

At present there is only one industrial method (the Envilack process) by which paint sludge can be reprocessed into new paint or base materials for the preparation of paint [199]. Another method (the Kluthe process) is now in development and has been tested on a bench scale [207].

The Envilack process (see Fig. 66) includes the following steps: dewatering, viscosity adjustment, purification, and separation into vehicle and pigment fractions. The process also permits reprocessing of paint sludge into new paint. If the conditions listed above are fulfilled, the following types of paint can be reprocessed:

1) Alkyd – aminoplast baking enamels
2) Acrylic resin paints and acrylate – aminoplast combinations
3) Polyester, epoxy, and melamine resin paints
4) Oxidizing drying paints

5) Physically drying paints

A 5000 t/a Envilack plant has been brought on stream in Duisburg, Germany. An expansion to 12 000 t/a is planned.

The Kluthe process processes paint sludge by distillation or extraction. The end products are a solvent phase, a vehicle fraction, and a pigment/filler fraction. This process can again be used only when the prerequisites cited above are fulfilled.

This vwk/GFR process [208] developed in 1990, yields an aqueous phase, an organic phase, and solids. This is less a paint sludge recycling process than a method of converting paint sludges to an easily handled form that facilitates recycling. A full-scale plant with a capacity of 6000 t/a has been put in service.

The process includes the following steps (see Fig. 67):

1) Drying of paint sludge with heat-transfer oil
2) Condensation of the gas phase from first step
3) Separation of the resulting liquid phase into organic and inorganic phases

The solids discharged from the process can be used in a variety of ways:

1) Auto underhood insulation, rear-window shelves, and other molded parts for the automobile industry
2) Insulating mats
3) Aggregate in cement production and brickmaking
4) Filler for plastics dispersions

Hydrogenation and Gasification. At Bottrop, VEBA Öl has operated a 200 000 t/a refinery residue processing plant since 1977 (see Fig. 68). About one-third of plant capacity is reserved for recycling special feedstocks, including distillation residues and paint sludges. The sludges must be inlet to the process in a dried, pneumatically conveyable form; there is currently no drying capacity at Bottrop.

The Bottrop plant can also hydrogenate plastics, thus making residues from resin-impregnating plants available for recycling even though relatively small quantities (a few t/a) make this an economically marginal proposition.

In addition to hydrogenation, organic residues (plastics, resins, etc.) can be processed in gasification systems like the one at Energiewerke Schwarze Pumpe. The same holds for resin wastes from impregnation plants, although small quantities make recycling more difficult in this case.

7.5. Metals

7.5.1. Introduction

Metal recycling can mean recycling of metal wastes to reclaim the metals, or recycling of nonmetallic compounds that contain metal cations to obtain the metals. This section deals only with compounds that were already in metallic form and became waste during production or consumption or in the course of recycling. Thus the winning of metals from primary resources and the wastes associated with these processes are not covered; the relevant waste sources in the product life cycle, which begins with the original shaping (e.g., casting or molding) of the metals, are considered.

Figure 69 shows the stages in metal recycling, based on VDI Guideline 2243 [209]. Each branch entails a decision whether a recycling path is available by which the waste can be recycled to a metal or other product in an economically and environmentally acceptable way and returned to production or consumption. The term metal recycling is used only when the result of recycling replaces a construction material or manufactured product derived from primary resources. Figure 69 does not reflect the fact that recycled substances and products can be forwarded to a process or use different from the initial one. According to the German Recycling Act (Kreislaufwirtschaftsgesetz) of July 8, 1994, which conforms to EC provisions, all process results that are not product must be viewed as wastes if they are not off-gases or wastewater. Wastes are classified as wastes for recycling and wastes for disposal (waste processing and landfill disposal). For example, production scrap becomes wastes for recycling.

In what follows, "recycling of production wastes" (Fig. 69, box 2) refers to wastes from the process chain, beginning with initial shaping and ending with fabrication of the finished product. In recycling during product use (box 4),

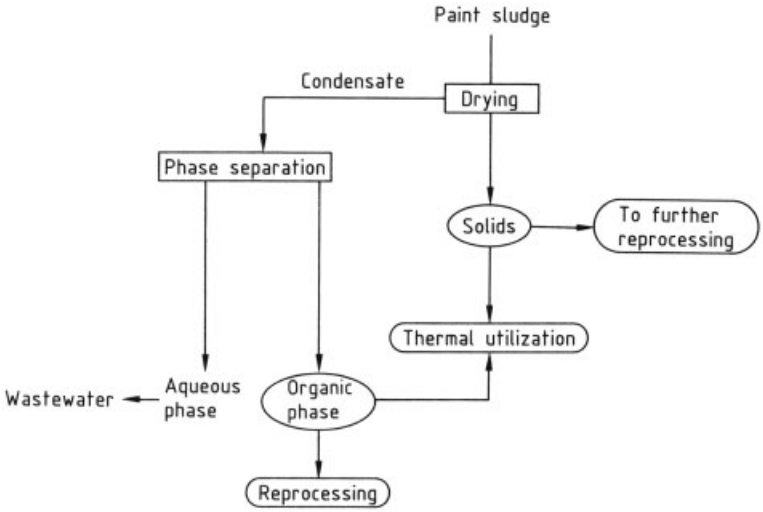

Figure 67. Flowsheet of the vwk/GFR process for processing paint sludge

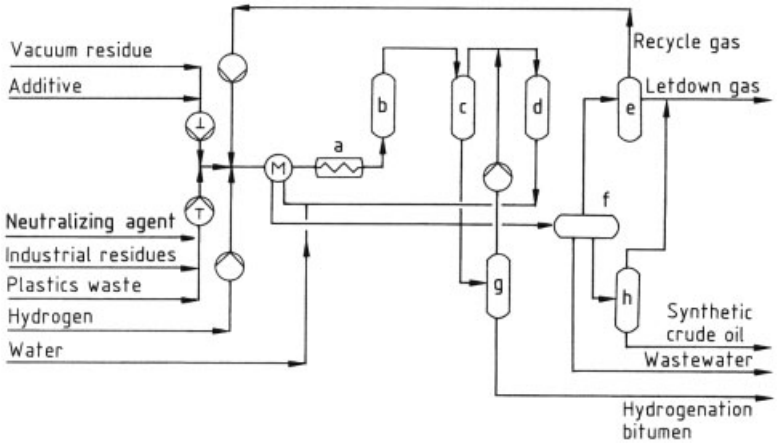

Figure 68. VEBA Öl hydrogenation process
a) Preheater; b) Liquid-phase reactor; c) High-temperature separator; d) Gas-phase reactor; e) Gas scrubbing; f) Low-temperature separator; g) Vacuum distillation; h) Stabilization

in contrast to all the other forms of recycling, the form of the product or component is not broken down. The product or component is put to new use after reconditioning (according to VDI 2243 [209]).

The recycling of product wastes is subdivided into product-specific recycling (box 6), material-specific recycling (box 8), and recycling of material wastes (box 10). In product-specific recycling, the processes are oriented to particular product groups or product- components, such as vehicles and batteries. Material-specific recycling and recycling of material wastes can be regarded as subclasses of product-specific recycling. In material-specific recycling, processes are adapted to a particular material or to particular material groups, such as the arc furnace for steel scrap and the salt-bath furnace for aluminum scrap. In recycling of material wastes, alloys and assemblies are subjected to metallurgical processes to recover their metals separately, as for example in secondary copper smelting.

While the feeds to the above recycling processes are generally metallic, nonmetallic wastes from the other processes are introduced into the recycling of residues (Fig. 69, box 12)

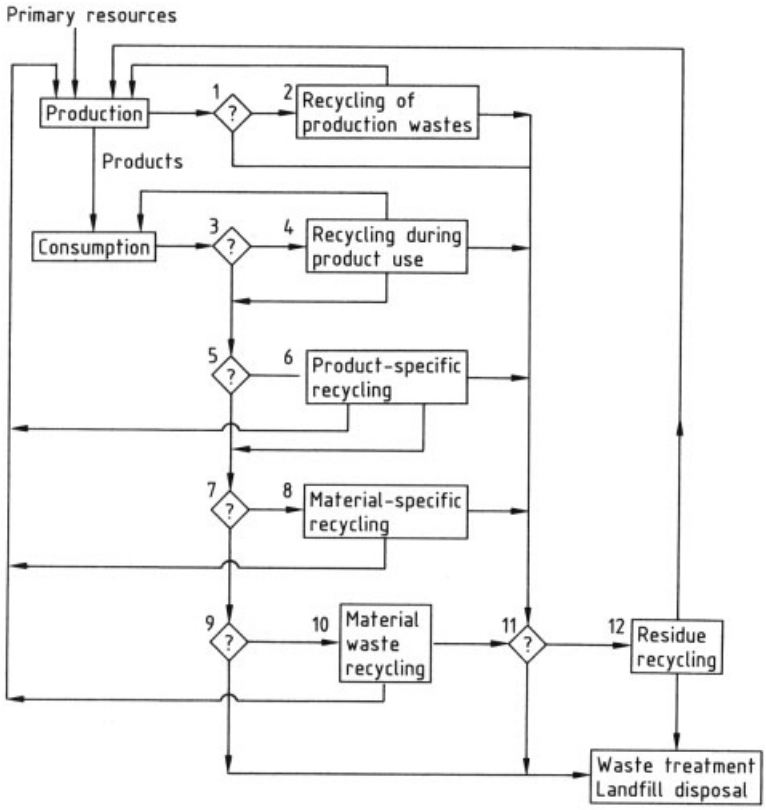

Figure 69. Routes for metal recycling

and refined to metals. Examples include grinding slurries from product finishing and salt slags and flue dusts from material-specific recycling and material waste recycling. Often residue recycling is integrated into the other recycling paths, so that the boundaries between forms of recycling are not always clear. Overlaps with residue recycling occur particularly in production waste recycling, where not only metallic but also nonmetallic wastes are treated.

The structure of metals makes it possible, in principle, to recycle them with no loss of quality. Recycling is subject to not only economic but also environmental and technological limitations. Environmental limits are encountered when more primary resources are consumed in and more environmental burdens result from recycling than are replaced by the reclaimed metals. Technological limits occur where process development is not yet justified on economic grounds.

Because of the large number of metals, their widely varying uses, and the resulting range of recycling routes, it is not feasible to treat metal recycling in a comprehensive way. Instead, significant recycling options for the economically most important metals based on waste products and wastes are presented.

7.5.2. Production Waste Recycling

In view of the restrictions stated in Section 7.5.1, the only criteria for classifying waste sources are manufacturing processes. Wastes are attributed initial shaping, forming, separation, and surface coating, and are also classified as new scrap, plant-return scrap (home scrap), or primary wastes.

Initial Shaping. Typical processes for the initial shaping of metals are the various forms of casting. Metal wastes from casting are crop ends

(lost head and base) in continuous ingot casting, gate and riser systems in shape casting, and aborts and metal samples. If these metal wastes are collected separately by composition, they can be returned directly to the melting process. Nonmetallic wastes generally arise through contact with atmospheric oxygen. Aluminum, a relatively reactive metal, forms a dross that must be conditioned in a special process (Section 7.5.2). If, however, the relatively noble metal copper absorbs oxygen, it need only be reduced by poling in the anode furnace, the final pyrometallurgical process before the manufacture of electrolytic copper.

Forming and Working. With regard to the wastes generated, forming is subdivided into hot and cold forming.

In hot forming processes such as hot rolling and forging, oxidation products (e.g., scale) are formed along with wasters, cuts, and samples. Scale, an oxidation product of iron, does not adhere to the steel and can be removed with high-pressure water during hot rolling. Metallic wastes are returned to melting, while scale is sintered and reduced in the blast furnace together with iron ore concentrates.

Wastes from cold forming processes (e.g., rolling, drawing, deep drawing, and bending) include cuts and specimens of not only single materials but also composites. Such a composite consists of several metallic materials or of metallic and nonmetallic materials. While metal wastes from single materials can be fed back to the respective melting operations, composites have to be subjected to a more differentiated production waste recycling. It must be decided whether all substances are to be recovered from the composite or not. Surface coatings are collected separately if they consist of a valuable metal or if they would interfere with the recovery of the parent metal. Silver, for example, occurs as thin coatings on materials used in the electrical and electronics industries. It can be removed from the substrate with sodium cyanide solution to which aromatic nitro compounds have been added as oxidizing agents. Copper and copper alloys are, however, attacked by the same agents. To recover the silver, the resulting solutions are subjected to ion exchange, electrolysis, or cementation, similarly to waste solutions from electrolysis [210].

To recover tin from tinplate waste containing 7 – 4.5 g of tin per kilogram of sheet steel, the plate composite is treated with gaseous chlorine in a reaction tower. The exothermic reaction yields tin tetrachloride $SnCl_4$, which is distilled off and condensed in a cooler. It can be used in making pink salt $(NH_4)_2SnCl_6$, organotin compounds, and other products. Tinplate scrap can also be electrolytically detinned by anodic dissolution in 5 – 10 % caustic soda solution, giving sodium stannate. Paint coatings are removed from the scrap with high-pressure water and hot soda solution; the cleaned scrap is then placed in movable, perforated steel balls or baskets. If the electrolyte made in this way contains Pb, Zn, As, or Sb, these are precipitated with sodium sulfide. Tin deposited on sheet-iron anodes from the electrolytes (50 g/L tin) is recovered by melting and may be refined again by steam injection. The detinned electrolyte (2 – 10 g/L tin) is returned to anodic dissolution [211].

Another way of recycling purely metallic composites so that none of the substances contained in them is lost is melting without prior separation. This is possible only in the case of material compatibility [212], i.e., that no element present in the wastes leads to a violation of the specification for the construction material produced by melting without dilution by primary metals. Tinned copper can be melted to obtain bronzes and red brass. A nickel- and chromium-plated steel can also be utilized in making chromium – nickel steels without prior separation. Melting must be carried out with very low oxygen partial pressures, since otherwise the chromium oxidizes and forms slag. Nonmetallic components are generally not recovered from composite materials.

Cutting leads to formation of punching waste, shearing scrap, and other materials. These correspond to wastes from cold-forming processes and are treated as such. Metal-removing operations yield turning, hobbing, milling, boring, or saw chips, depending on the technique and the geometrical form of the tool used. These operations are done with coolants, which consist of ca. 10 % mineral oils, water, and a variety of additives. Before the chips can be inlet to a melting process, they must be de-oiled to comply with emission limits for organic substances. Oil can be removed by centrifugation and wash-

ing. Grinding slurries are processed when they contain valuable metals or compounds of such metals.

The U.S. Bureau of Mines has developed a hydrometallurgical process for recovering cobalt, chromium, manganese, nickel, and tungsten from grinding slurries generated in the machining of sintered metal alloys [213]. A slurry containing (after drying and de-oiling) 15.1 % cobalt, 19.8 % chromium, 31 % iron, 3.6 % manganese, 11.5 % nickel, and 5.1 % tungsten is produced in a quantity of about 1600 t/a in the United States. The de-oiled and washed slurry is leached in hydrochloric acid with addition of gaseous chlorine. The solution contains Co, Cr, Fe, Mn, and Ni; the residue contains tungsten. From the leach residue, Na_2WO_4 is obtained by dissolution in caustic soda, then calcium tungstate $CaWO_4$ by precipitation with calcium chloride. From the first leach liquor, iron, cobalt, manganese, nickel, and chromium are obtained in succession by solvent extraction of their chlorides. The residual solution can be reused for leaching. All products have impurity levels below the specified levels and are thus salable. In the laboratory-scale plant described, 90 % of the tungsten and over 97 % of all other values were recovered [213].

The grinding of permanent magnets made from iron alloys containing aluminum, nickel, and cobalt yields residues containing 10 – 15 % cobalt and 4 – 7 % nickel. Sulfur and carbon present in the grinding slurries interfere with the recovery of these valuable metals. The sulfur originates in the binder of the abrasive, the carbon from the coolants used. These substances must be removed before the valuable metals in the grinding slurries can be reused directly for permanent-magnet production. A process devised by the U.S. Bureau of Mines consists of desulfurizing and de-oiling at 112 °C with tetrachloroethylene, which is recycled, drying and grinding the residue, screening and magnetic separation to remove the main component of the abrasive (Al_2O_3), oxidative roasting at 950 °C to remove residual sulfur and carbon, and reduction of the metal oxides at 1200 °C with hydrogen [214]. The result is a metal powder that contains > 95 % of the cobalt, nickel, and iron, as well as ca. 90 % of the copper fed to the process. The contents of residual sulfur and carbon in the metal powder are 0.01 % and 0.06 % respectively. This material can be added directly to the melt for the fabrication of permanent magnets; the ca. 30 % Al_2O_3 in the recycled product is slagged on melting [214].

Pickling solutions are generated when metals are etched or cleaned with acids such as hydrochloric acid. The consumption of hydrochloric acid for this process (as commercial 37 wt % acid) is 0.3 – 3 kg/t steel; without regeneration it would be 20 – 35 kg/t steel [215]. The iron(II)-loaded pickling solution is treated with HCl gas, whereby $FeCl_2 \cdot 4\,H_2O$ crystallizes out and is separated. The solution is recycled to the pickling process. The iron chloride is heated to ca. 300 °C in a stream of oxygen and steam, giving Fe_2O_3 and HCl. This HCl gas is used to treat iron-laden pickling solution to crystallize iron chloride. The hydrochloric acid thus moves in a closed circuit [216]. The ca. 4.8 kg iron oxide generated per tonne of steel treated is used as a raw material in other processes [215].

Coating. In electrolytic operations, process water is generated by rinsing and the disposal of short-lived process baths; this water has to be purified. The addition of milk of lime or caustic soda, along with flocculation aids if necessary, converts the dissolved heavy metals to sparingly soluble hydroxides or hydrated oxides, which precipitate to form "cell mud," which is then mechanically dewatered. In the form of hydroxides or hydrated oxides, the mud contains metals that are to be electrolytically deposited on the workpiece. Table 12 lists average metal levels in U.S. electrolysis cell muds.

These muds can be treated by hydrometallurgical or pyrometallurgical methods. Standard hydrometallurgical processes include leaching, solid – liquid separation, cementation, and precipitation. Chlorides, fluorides, phosphates, ammonium, and organic substances interfere with the processes. Even though hydrometallurgical processing of cell muds exhibits high flexibility, the small amounts of muds (<100 t/a for most electrolytic processes) and the variations in mud composition, depending on the electrolytic operation, make these processes very cost intensive [217]. Pyrometallurgical processes are less flexible but have cost advantages. The pyrometallurgical processes used for treating cell muds are the same ones developed for the refining of primary resources or in some cases for other

Table 12. Average analyses (wt %) of electrolysis cell muds in the United States (1991), arranged according to dominant metal [216]

Number of samples	Cu	Ni	Pb	Zn	Cr	Fe	Moisture
171	16.4	7.5	0.6	3.6	1.4	4.7	43.3
148	2.8	16.1	0.1	1.8	3.3	3.8	41.4
33	3.7	0.1	36.0	6.4	0.5	1.1	16.5
99	2.9	0.4	3.7	28.8	1.6	4.8	31.7
57	1.1	1.7	4.8	1.8	24.4	3.5	34.7

wastes. Such plants have high capacities, in the range of 50 000 – 300 000 t/a, which is at least a thousandfold greater than the output of cell mud from a plant. Any metallurgical process is suitable only for muds having the combination of metals for which the process was first developed and is now operated. With express approval for the use of cell mud as a secondary raw material, the following pyrometallurgical facilities are available [218]:

1) For muds containing cadmium, lead, zinc, and iron: the B.U.S. rotary-furnace volatilizing process at Duisburg [219]
2) For muds containing nickel, chromium, and iron: the Swedish Plasmadust process [220]
3) For muds containing zinc: the New Jersey process of Harzer Zink GmbH at Goslar [221]
4) For muds containing copper and nickel: the secondary copper process operated at Lünen by Hüttenwerke Kaiser AG (see Section 7.5.4.3) [222]

7.5.3. Recycling during Product Use and Product Reconditioning

Consumer and capital goods are taken out of use because they are defective, because they no longer fulfill their function properly, or because they no longer correspond to the state of the art. Such a waste product must be examined (Fig. 69, box 3) to see whether reconditioning (box 4) is an option. In reconditioning, the form of the product or component is not broken down. Often not the entire product is reconditioned, only components of it, such as automobile engines or the roll sets from hot- and cold-strip rolling mills. One then speaks of making replacement products [223]; this is a subset of product reconditioning. The following steps are involved:

1) Completely disassemble item
2) Clean parts
3) Inspect and sort parts
4) Recondition or replace unusable parts
5) Reassemble item

Products reconditioned in this way are of equal quality to new products. Because the product form is preserved and products/components/parts are only partially reworked, they can be kept in service at a relatively low cost. The focus in component reconditioning is on machining operations such as turning, milling, boring, and grinding. This work makes up some 45 % of total reconditioning effort for a gasoline engine from an automobile [223].

The metal wastes generated in this process can be recycled as production wastes (Fig. 69, box 2) if they are unmixed and of known composition, and as wastes otherwise (box 10). Parts that can no longer be reconditioned pass on to product waste recycling (boxes 6, 8, and 10). Other examples of product reconditioning include household appliances such as water heaters, all kinds of power tools, office equipment such as typewriters and copiers, and automatic dispensers for cigarettes, beverages, and so on. Reconditioning applies particularly to metallic materials, since these can be machined without problems, restored to specified dimensions by buildup welding, and heat-treated or otherwise treated to impart the required properties. Partial disassembly of automobiles allows a 15 % reuse of components (motors, transmissions, generators, etc.); the reuse figure is 5 – 10 % for computer equipment [224].

7.5.4. Product Waste Recycling

Product waste recycling (Fig. 69, boxes 6, 8, and 10) includes product-specific and material-specific recycling as well as recycling of material

wastes. In contrast to recycling during product use (reconditioning), here the form of the old product, component, or part is broken down.

7.5.4.1. Product-Specific Recycling

In product-specific recycling, the recycling route is oriented to a particular product or product group. Product-specific recycling can be classified in a way that corresponds to the distinction between (discrete) manufacturing and (continuous) processing [224]. Discrete operations relate to disassembly (DIN 8580) [225], while all continuous operations in this context involve changes in the composition, properties, and nature of the substances. The continuous processing class is once again broken down into physical and chemical operations. As a rule, discrete operations precede continuous ones. In partial disassembly, components that are still functional or can be reconditioned are removed, as are those that would interfere with continuous processing and those that contain materials reclaimable after separation. In the case of an automobile, examples in the above order would be the alternator, the gasoline engine, and copperbased components. Table 13 lists the fractions of some consumer products that can be reconditioned and reclaimed, together with future target reclamation figures for Germany. The sums of the component reconditioning and reclamation fractions correspond roughly to the metal contents of these consumer products.

Table 13. Component reconditioning and reclamation of substances from used consumer goods (in wt %) [223]

Used consumer goods	Component reconditioning	Reclamation	
	1990	1990	Future
Automobiles	15	60.5	77.0
Washing machines		67.1	71.7
Dishwashers		58.0	88.0
Refrigerators		56.5	96.9
Computers	5 – 10	81.3	94.7
Televisions		5.0	83.0

7.5.4.1.1. Automobiles

The automobile is a highly complex consumer product. When it is taken out of service, it consists of:

1) Components that can be reused without reconditioning
2) Components that can be returned to use after reconditioning
3) Substances whose value justifies reclaiming them
4) Substances that interfere with reclamation

The average composition of a discarded 1993 auto with a total weight of 1126 kg [225], [226] was:

Steel	513.0 kg
Cast iron	104.5 kg
Aluminum	50.0 kg
Lead	10.2 kg
Copper	9.5 kg
Zinc	4.7 kg
Nickel	0.3 kg
Other nonferrous metals	32.3 kg
Plastics	248.0 kg
Glass	38.0 kg
Textiles	15.2 kg
Rubber from used tires	40.0 kg
Others	60.3 kg

These cars consist of 64.3 % metallic materials and 22 % plastics. Recent years have seen a marked reduction in the content of ferrous metals and a rise in the contents of aluminum and plastics.

Figure 70 shows the principal routes for auto recycling. Removal of fuel and fluids is followed by partial disassembly. Components such as the engine, alternator, starter, transmission, and so forth are either used as spare or replacement parts or reconditioned for reuse. As a rule, the lead – acid battery and the catalytic converter are also removed. Recycling practices for lead-acid batteries have long been established (see Section 7.5.4.1.3); the catalytic converter is of interest because it contains 1.7 – 1.9 g of platinum-group metals (see Section 7.5.4.1.2).

One purpose of automobile recycling is to reclaim the metals. The choice of method is governed not only by economics but also by the separation of heavy metals from ferrous metals. Heavy metals can impair both the service properties and the processability of steel, especially if several heavy metals are simultaneously present. Copper levels of 1000 – 1500 ppm should be aimed at, though high-tonnage steels can tolerate higher levels. For deep-drawing steels that contain tin and copper, the copper

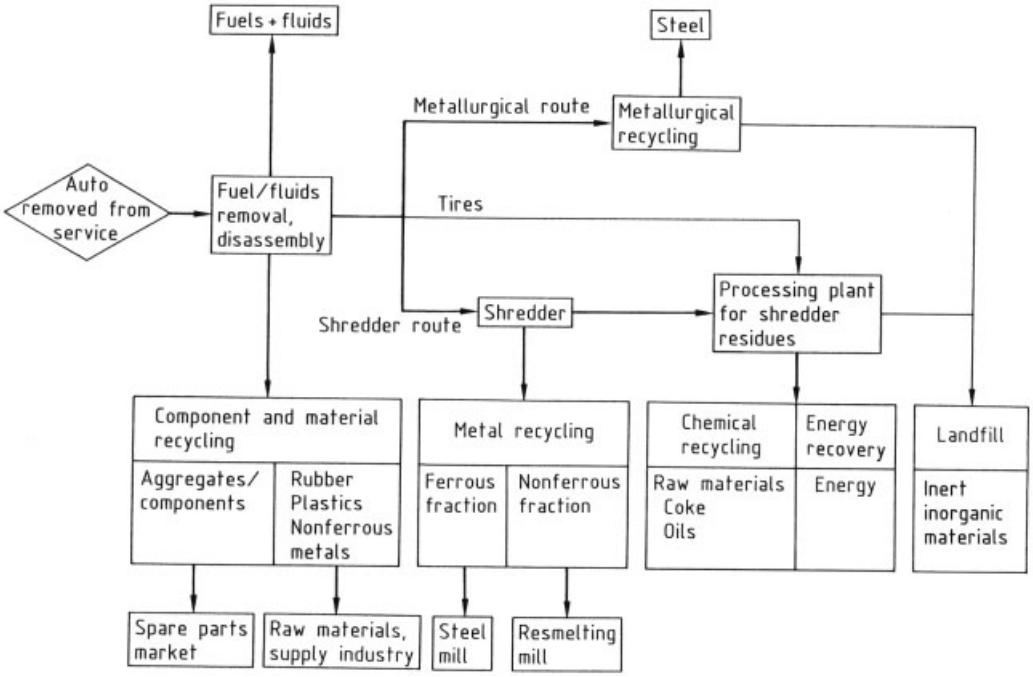

Figure 70. Conceptual elements and material streams in automobile recycling [226], [228]

content must be still lower. Otherwise, thermally stable copper – tin compounds can form, which, unlike copper alone, do not redissolve in the iron after precipitation at grain boundaries. Studies on deep-drawing steels with 150 – 650 ppm copper have shown that the key mechanical properties (yield strength, tensile strength, reduction of area at failure, ultimate elongation) after hot rolling did not change as the tin content was raised stepwise from 80 to 250 ppm. Nor could alterations be detected in the cold-rolled strip [227].

Heavy metals such as zinc and lead, as well as tin in scrap iron, are vaporized at the temperatures prevailing in the oxygen converter or arc furnace and also in metallurgical recycling. Complete removal of heavy metals is not possible, however, because the activities of these metals also decline with decreasing concentrations in the steel melt. The majority of copper inlet with scrap remains in the steel together with ca. 60 % of tin. For this reason, care must be taken – especially in auto recycling – to remove copper and tin before metallurgical recycling. Separation and reclamation of the metals from autos is a substantial contribution to resource conservation. The removal of nonferrous metals from the auto can be performed by discrete assembly (or partial disassembly) operations or by continuous processing of shredder fractions. Of the ca. 2.6×10^6 autos discarded in Germany in 1992, ca. 85 % were shredded and ca. 15 % cut up in scrap shears or compressed in baling presses.

Shredder Route. After partial disassembly and removal of fluids, cars are mechanically reduced in size by the shredder. The selective action of the shredder disintegrates complete car bodies; ferrous components are compressed to fist-sized pieces. Components entrained in the exhaust air are removed in an integrated cyclone; this shredder light fraction has been disposed of in landfills, but studies are under way on reclamation of valuable constituents from the light fraction as well as extraction of the energy content as heat. The heavy fraction, 77 wt % of the inlet stream, is sorted on the basis of several physical properties. Besides magnetic separation to isolate ferrous scrap and pneumatic separation to recover nonmetals, other processes are becoming established. An example is eddy-current separation used to obtain nonferrous

metal concentrates. Methods are being developed for automatic identification and separation of metal mixtures by composition. The feed material must be largely free of interfering substances and must be disintegrated to the point that piece sorting is equivalent to material sorting (see Section 7.5.4.2).

The average mass balance for old cars put through shredders in the Federal Republic of Germany in 1990 shows that 70 wt % of the feed goes into the scrap-iron fraction, 23 wt % into the light fraction and 7 wt % into the mixed fraction consisting primarily of nonferrous metals. Figure 71 shows a breakdown of the mixed fraction (into aluminum, copper, zinc, and lead) for the same period. About 1 % of aluminum and 10 % of copper remain in the iron fraction, and this copper contributes to copper levels in the steel. Zinc is not present in the ferrous fraction, suggesting that the bodywork in question was mainly nongalvanized. The light fraction and iron fraction data give the losses as fractions of shredder feed. Under the mixed fraction, the fractions processable in metal foundries or smelters are indicated. These data may vary with age of the automobile and extent of disassembly.

Metallurgical recycling, developed and put through pilot tests by Voest-Alpine-Stahl and Mercedes Benz, is a novel method for processing automobiles after service. As in all other recycling approaches, the first step is removal of fuel and fluids. This is followed by partial disassembly aimed at maximum recovery of spare parts (i.e., reuse of assemblies and components) and the generation of fractions suitable for materials recycling. This partial disassembly step also determines the copper content of the steel to be made. An analysis of a middle- class 1980-model car shows that ca. 74 % of total copper is in the engine with accessories, the transmission, and the main electrical system. If the engine, transmission, starter, alternator, radiator, and 50 % of the main electrical wiring are removed, the remaining scrap steel contains ca. 0.2 % copper [228]. The time required to remove these components is ca. 25 min, and their removal lowers the total weight of the car by ca. 47 %. The scrap, still contaminated with about 100 kg of plastics and organic constituents, is baled and fed to the scrap melter. The melter is an enclosed shaft furnace, charged continuously and heated with natural gas/oxygen burners. Adhering plastics lower the energy requirement for melting by 30 – 40 %. In the high-temperature zone, organic substances are oxidized and go into the gas phase. Zinc, lead, and other heavy metals are vaporized and collected in a filter. In the upper part of the reactor, at temperatures above 1100 °C, dioxins and furans are destroyed. The lead – zinc dust obtained from gas cleaning is utilized in zinc smelters (Section 7.5.5). The steel produced in this way has a low carbon content and does not differ in quality from steels made by the shredderroute. This process has the advantage that for ca. 530 kg of material inlet to metallurgical recycling, 143 kg of light fraction does not have to be disposed of in the landfill as in the shredder process. Utilization of the energy content of residual organic substances results in a saving of primary energy, while the presence of residual inorganic components means that slag formers need not be added to the melt. The process has the disadvantage that the aluminum from the auto is lost as slag [226].

Complete Disassembly. The aim of disassembly is the separate and complete collection of components and substances that can be reclaimed or that must be classified as hazardous wastes, with appropriate interim storage, reclamation, or treatment of the collected materials. Disassembly can be carried out in either of two ways: line disassembly and island disassembly. Disassembly organized similarly to assembly is possible only in plants that are restricted to certain models of vehicles. Because car models are so diverse and are not yet designed for disassembly, this operation cannot be completely automated; it is therefore set up in open facilities with semiautomated conveyor lines. In disassembly, an automobile consisting of some 10 000 parts is broken down so that they can be sorted into about 10 – 15 pure material classes. When disassembly is done on the island principle, all work is performed in disassembly cells or islands. The disassembly line approach has the operations performed in sequence and at different locations; in this way they can be optimized and high capacities can be achieved. Island disassembly permits greater "depths" of disassembly. After disassembly, the body is baled in a press to the density of scrap iron for economical

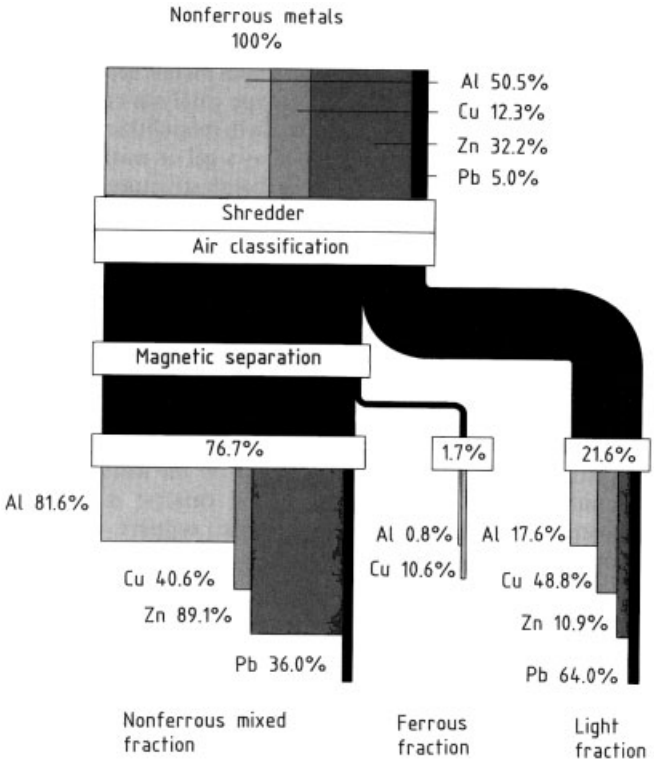

Figure 71. Distribution of nonferrous metals among shredder fractions [226]

shipping. Preliminary removal of all nonferrous metals makes it possible to lower copper and tin levels in the scrap and thus improve the quality. The principal nonferrous metals present are aluminum from wheel rims, cylinder heads, and oil pans; zinc from a variety of die-cast parts; copper from engine parts, starters, generators, radiators, and cables; and lead from batteries and wheel-balancing weights.

Sheet-Metal Recycling. Contamination from surface coatings and automotive fluids enter the steel melt when auto scrap is fed to secondary steelmaking. Heavy metals and organic compounds must be retained by expensive filtration. A method has been developed by which body parts can be directly reused as sheet metal [229]. Large pieces of sheet metal, provided they are not damaged by rust or accidents, are removed, heated to remove coatings, and then straightened. This means that the steps of shredding, casting, rolling, descaling, and recrystallization annealing can be omitted. In comparison with steel-mill processing, the process yields an energy saving of ca. 97 %. The properties of the sheet metal are comparable with those of new metal.

7.5.4.1.2. Catalytic Converters

In 1987, automotive catalytic converters represented a demand for 35 t of platinum (out of a total consumption of 102 t), 7.3 t of rhodium (out of 10 t), and 6.4 t of palladium (out of 101 t) [230]. Consumption will rise, since virtually all used cars will have catalytic converters in the foreseeable future. Because of their average contents of 1.5 g platinum and 0.3 g rhodium the converters are removed from discarded autos before further processing [226]. Catalysts used in the United States also contain palladium. The ratio of platinum to palladium to rhodium is about 14: 7: 1. The platinum metals are supported on γ-Al_2O_3. Pellet-type catalysts consist entirely of γ-alumina, while in monolithic catalysts this material is applied as a gel or wash coat on the

surface of the honeycomb structure, made of iron cordierite $2\,FeO \cdot 2\,Al_2O_3 \cdot 5\,SiO_2$ or magnesium cordierite $2\,MgO \cdot 2\,Al_2O_3 \cdot 5\,SiO_2$. The pellets in a pellet catalyst weigh 2 kg on average; a monolith averages 1 kg with about 15 % γ-Al_2O_3. Both types have the same content of platinum group metals, so that the metal concentration in monoliths is about twice as high as in pellets [231]. The share of pellet catalysts in exhaust converters is declining.

The first step is to cut off the stainless steel case in which the actual catalyst is located, then remove the ceramic support material. The support is ground, homogenized, and analyzed to determine the noble metal content. Smelters or parting plants recover the platinum group metals by hydrometallurgical or pyrometallurgical methods. In hydrometallurgical processes, sulfuric acid is used to dissolve the material containing the platinum group metals: either the complete pellets or just the alumina coat from the cordierite monolith. Various combinations of separation and enrichment methods are employed [231]. In a method applicable to both types of catalyst the ground material is leached for 1 h at 160 °C with 1 % sodium cyanide solution. The dissolved platinum metals are precipitated by decomposing the cyanide complexes at 260 °C. Any lead present is not dissolved; it enters the leach residue [232]. The wet chemical processes are problematical because of the large amounts of wastewater and residues involved. The yield of platinum group metals is lower than that from pyrometallurgical processes.

In pyrometallurgical recycling processes, the catalysts are melted with a collector metal. Degussa has a lead blast furnace in which the platinum metals together with noble metals from other residues are collected in lead. The ceramic material is slagged, and the noble metal free slag is discarded. The resulting lead bullion is forwarded to the converter, where it is oxidized in the melt by oxygen or air. Oxides of non-noble metals dissolve in the litharge produced in this cuppelation step, leaving behind auriferous silver, an alloy of silver, gold, and the platinum group metals. The individual noble metals are won from this metal by standard methods [233].

If copper is used as collector, the catalysts are inlet to a standard copper recovery process. The noble metals accumulate in the anode slime of the electrolytic copper refining cell and are extracted from this in conventional fashion [233]. It is claimed that the highest recycling rates for platinum group metals are achieved with this technique, but the time from when the catalysts enter the process until the platinum group metals are obtained is much longer than in other recycling methods [231].

The Tetronics plasma furnace melts the catalysts with the addition of iron at temperatures above 2000 °C. Part of the silicon dioxide is reduced to silicon, so that the platinum group metals are collected in an iron – silicon alloy. The pulverized metal is treated with sulfuric acid under oxidizing conditions to dissolve the iron and silicon. The noble-metal concentrate in the residue is refined to the metals in a parting plant.

The recycling rate is in the range of 86 – 93 % for platinum and palladium, and between 81 and 87 % for rhodium [232]. Both the energy consumption and the recycling rate are lower in the hydrometallurgical processes than in the pyrometallurgical processes.

7.5.4.1.3. Lead – Acid Batteries

On average, automobile batteries weigh 16.5 kg and contain ca. 10 kg of lead (ca. 60 %). More than 50 % of the lead consists of sulfate and oxide paste. Of 78×10^6 lead – acid batteries in the United States in 1991, more than 98 % were recycled; the rate for Europe in the same period was ca. 90 % [232].

Figure 72 shows a flowsheet for mechanical separation of lead – acid batteries as practiced by Harz-Metall-GmbH, an operation of Metaleurop. The battery cases, made of polypropylene or, to a decreasing extent, hard rubber (ebonite), are mechanically broken up in the double-roll crusher. The sulfuric acid released is neutralized with milk of lime. By grinding, screening, and heavy-medium separation, the battery scrap is divided into metal, oxide, polypropylene, and ebonite fractions. Of these four, the ebonite is not processed further at present [234].

At Harz-Metall, the metal and oxide fractions are obtained in roughly a 1: 1 ratio; polypropylene accounts for about 4.5 % based on battery scrap. The oxide fraction contains 69 – 75 % lead, 0.3 – 0.5 % antimony, and 5.3 – 6.2 % sulfur, while the metal fraction contains 60 – 75 % metallic lead and 10 – 30 % lead in oxide form.

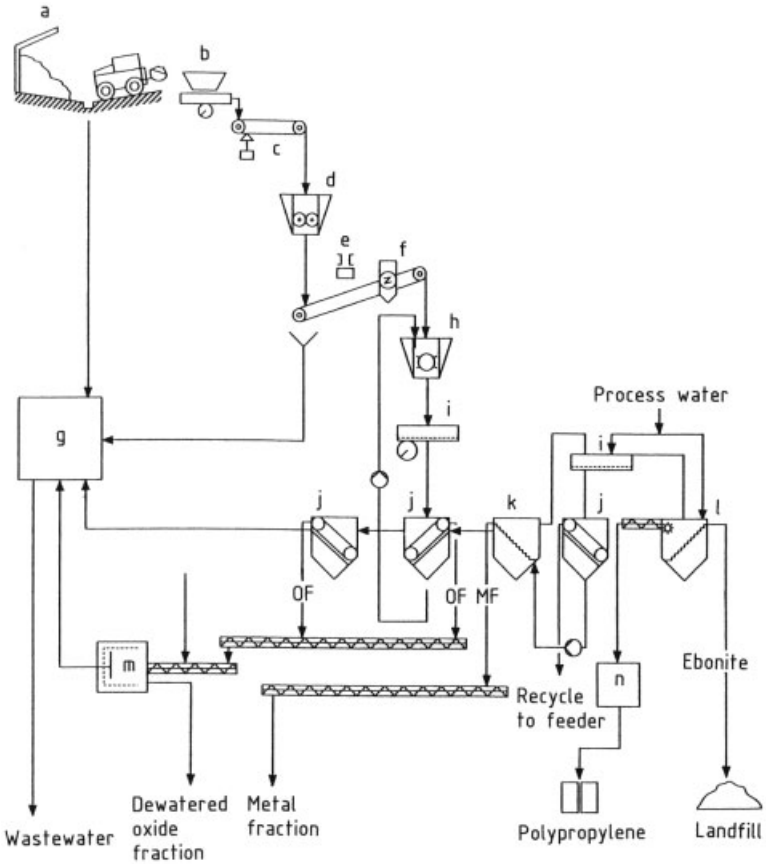

Figure 72. Flowsheet for physical processing of lead – acid battery scrap [234]
MF = metal fraction, OF = oxide fraction
a) Battery scrap storage; b) Feeder; c) Conveyor scale; d) Double-roll crusher; e) Magnet; f) Metal detector; g) Wastewater treatment; h) Hammer mill; i) Vibrating screen; j) Oxide classifiers; k) Hydraulic classifier; l) Plastics classifier; m) Pusher centrifuge; n) Polypropylene refining

The lead losses of < 1 % result from the 4 – 11 % lead content of the ebonite fraction. The oxide fraction is inlet to the rotary furnace; coal is added for reduction and iron filings and soda to fix the sulfur. The furnace product is lead containing 1 – 2 % antimony. The metal fraction is processed in the short rotary furnace, yielding lead with 3 – 4 % antimony. The low-antimony lead is refined to soft lead (99.97 % Pb), the high-antimony product to hard lead [233]. The difference between these secondary leads and primary lead (99.99 % Pb) is that bismuth and silver are not removed by the refining that takes place in recycling smelters [235].

In contrast to the above fraction-by-fraction process, the blast-furnace process (RSR in the United States, VARTA in Germany, Bergsøe in Denmark) dispenses with mechanical crushing into fractions. The battery case is automatically drained, so that whole, unground batteries can be charged to the blast furnace together with coke, limestone, and iron. The coke releases energy and establishes a strongly reducing atmosphere in the furnace. Organic components of the battery – the case material, spacers, and seals – contribute to the reducing atmosphere as well. Limestone and iron form a fluid slag with the inorganic constituents of the battery and the coke. Part of the iron, together with reduced sulfate sulfur, forms an iron – lead matte containing ca. 8 % lead. Nearly 90 % of the inlet sulfur is fixed in this way; the matte is contin-

uously discharged with the slag and separates from it because of the density difference. After solidification, the slag and the iron – lead matte can be mechanically separated. The top gas contains carbon monoxide, chlorides, and decomposition products of the plastics. It is combusted at 900 –1000 °C with a slight air excess, cooled, and passed through dust- collection equipment [236].

Plant used for recovering lead from batteries in the United States includes blast furnaces and, increasingly, reverberatory furnaces. The lead-bearing components of the batteries are processed in the reverberatory furnace with limestone, iron, and coke; the furnace discharges soft (low-antimony) lead bullion and an iron – sulfur matte. The slag is rich in lead oxide and contains most of the inlet alloy constituents; it is processed in the blast furnace or electric furnace to obtain hard lead and a slag that can be disposed of in landfills [232], [235].

7.5.4.1.4. Dry-Cell Batteries

Primary batteries are not rechargeable, in contrast to secondary or storage batteries. Various types of cells have anodes of zinc or lithium; cathodes of manganese dioxide, silver oxide, mercury oxide, or carbon; electrolytes of ammonium or zinc chloride, potassium hydroxide, or organic compounds; and cases of cardboard, plastics, and primarily steel. 1000 kg of such batteries, as collected, contain the following substances [237]:

Manganese dioxide	270 kg
Iron	210 kg
Zinc	160 kg
Graphite	60 kg
Ammonium chloride	35 kg
Copper	20 kg
Potassium hydroxide	10 kg
Mercury	3 kg
Cadmium	0.5 kg
Silver	0.3 kg

The composition varies with the proportions of different battery systems in use and due to fluctuations in cell composition. For example, the consumption of mercury for zinc – carbon and alkali – manganese cells fell from 30 t/a in 1986 to 5.5 t/a in 1988 [238]. In the Federal Republic of Germany, some 18 000 t of dry- cell batteries were sold in 1988; roughly one-third of these were collected for recycling.

The inhomogeneous mixture of dry batteries is subjected to thermal treatment either unsorted (Sumitomo and Recytec processes) or after preliminary sorting into button and cylinder types (Voest-Alpine and Clean Japan Centre processes). In the Recytec process, water, organic substances, and mercury are distilled and condensed, and the off-gas stream is cleaned with adsorption filters. After the solid residues are ground, manganese dioxide and zinc are leached out and separated by electrolysis. The ferrous components are removed from the insoluble residual fraction by magnetic separation. The remaining nonferrous metals are selectively dissolved with tetrafluoroboric acid and recovered separately in several electrolysis steps (zinc – cadmium, iron – chromium – nickel, copper – silver). In the Sumitomo process, fractional distillation is used to remove mercury (500 °C) and organics (800 °C), then the solid residues are fed to a blast furnace, where they are melted under reducing conditions at 1350 – 1500 °C. Zinc vaporizes and is condensed as the metal. Iron and manganese are discharged as an alloy that can be used in making ferromanganese or silicomanganese. A multistage off-gas treatment system recovers mercury and also ensures very low emission levels.

In the Voest-Alpine process, button cells are heated to 650 °C under reducing conditions to drive off mercury, which is condensed in metallic form. The residue is leached with nitric acid. The stainless steel scrap recovered after this step is forwarded to scrap storage for remelting. Treatment of the solution with hydrochloric acid precipitates silver chloride, which is reduced with zinc metal to obtain silver. In this way, silver, mercury, and stainless steel scrap, all of salable quality, are reclaimed from button cells. The wastes are zinc chloride and zinc and nickel slimes. Cylindrical batteries are ground in the Voest-Alpine process, the paper and plastics parts are removed by screening, and the scrap iron is recovered by magnetic separation. From the residue, mercury and zinc are successively vaporized and recovered as the metal or oxide. The residual mass can be converted to electrolyte (MnO_2).

7.5.4.1.5. Electronics Scrap; Circuit Boards

Because no unambiguous definition exists for electronics scrap, so the presence of circuit boards can be regarded as a first characteristic of this kind of scrap. Because electrical equipment also contains circuit boards, the broadest definition term would include such equipment along with computer, communications, and entertainment electronics. In Germany, 130×10^6 electrical and electronic devices with a mass of around 1.5×10^6 t are retired yearly. These old devices contain 30 000 – 45 000 t of circuit boards [240]. The average composition of electronics scrap in the United States (computers, telephones, relays, etc.) is reported as 30 % plastics, 30 % refractory oxides (15 % SiO_2, 6 % Al_2O_3), 40 % metals (20 % copper, 8 % iron, 4 % tin, 2 % each nickel, lead, and aluminum, and 1.2 % zinc), and 0.3 % noble metals [241]. The scrap analysis varies widely, depending on the components present and where they were manufactured.

The purpose of processing electronics scrap is to reduce pollutant levels (e.g., to remove capacitors containing polychlorinated biphenyls for underground deposition) and to recover metals, especially noble metals. The various routes for electronics scrap processing begin with preparation for hydrometallurgical or pyrometallurgical processing. The discarded products are manually dismantled into enclosures, motors, transformers, relays, circuit boards, cables, picture tubes, and other components; they are sorted into reusable components (transformers, memory chips, tantalum capacitors, etc.), reclaimable substance groups (noble metals, nonferrous metals, glass), and pollutant contents (PCB- containing capacitors, mercury/cadmium- containing primary and storage batteries). Recycling processes for circuit boards are presented in the form of a Sankey diagram in Figure 73 [242]. Mechanical grinding of the several sorted fractions follows in hammer mills, rotary knife cutters, or ball mills; the discharge particle size is such as to permit separation or reprocessing of the substances. Substances are separated according to their physical properties: by magnetic separation, eddy-current separation, dense-medium separation, and air classification. The resulting copper, aluminum, and iron fractions can be forwarded to appropriate industrial refiners or smelters.

Components that contain noble metals (circuit boards, contacts, etc.) are processed separately. To embrittle plastics for better grinding, the boards are sometimes heated to 425 °C. Hammer milling, especially for ceramic substrates, is followed by particle-size reduction in a ball mill. Without separation of nonmetallic and metallic substances, which is associated with losses of noble metals, the material obtained is forwarded to a copper or nickel smelter for further processing.

Hydrometallurgical processes must be considered for small operations where quicker recovery of noble metals is desired. Gold films can be removed from various substrates (along with copper, silver, and platinum group metals) by leaching with sodium cyanide and an oxidizing agent at room temperature. Gold, copper, silver, and the platinum metals are cemented with zinc, and the cementate is washed, dewatered, and melted in an induction furnace. The solution is regenerated and reused. The product metal contains 60 – 85 % gold, and the recovery rate is over 99 % [241].

Dissolution of metals from circuit boards with peroxodisulfates also decomposes pollutants such as polybrominated diphenyl ether flame retardants. Peroxodisulfates have an oxidation potential (+2.06 V) comparable with that of ozone. As Figure 74 shows for the example of copper, they can be anodically regenerated by electrolytic depletion of metals and returned to the circuit [243].

In a pyrometallurgical process, plastics are removed from the mechanically processed circuit boards removed by pyrolysis and the gas is combusted with an excess of oxygen. The solid residue can be melted and forwarded to a copper smelter for refining, or directly processed. The dust or sludge separated from the off-gas contains not only tin and lead but also noble metals, and must be subjected to further processing [241]. The solid residue, if directly processed in the recycling plant, must not contain any organic carbon, which would lead to excessively viscous slags in the arc or plasma furnace. A fluid slag is necessary for pyrometallurgical enrichment of noble metals in the metal phase. The slag system chosen depends on whether iron or copper is used as the metal collector. The slag system for iron is $CaO - SiO_2 - Al_2O_3$ and for copper, $FeO - SiO_2 - CaO$. If copper is used as

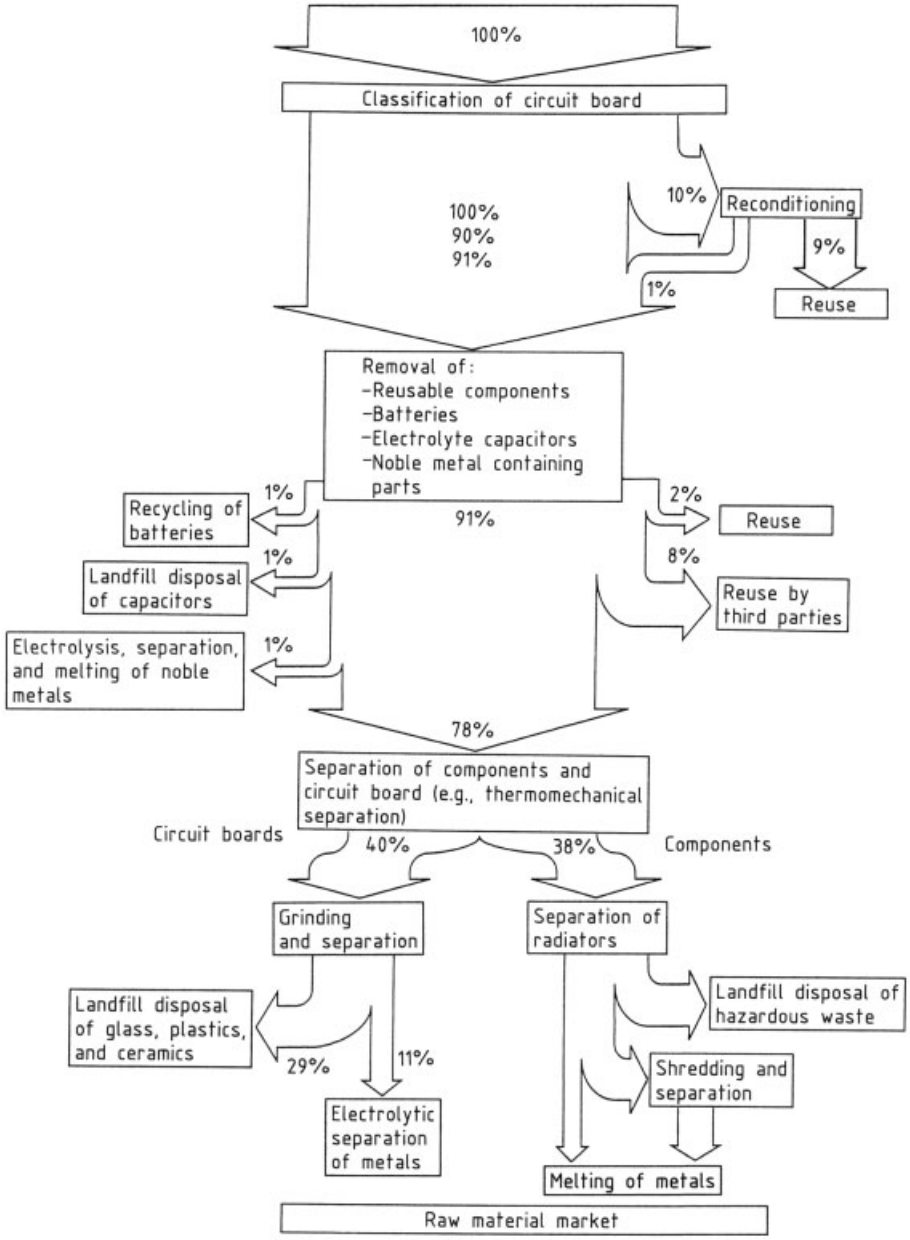

Figure 73. Sankey diagram for recycling of printed circuit boards [242]

collector, the process can be carried at 1350 °C and the noble metal containing copper can be cast into anodes. The noble metals accumulate in the anode slime during electrolytic refinement and are recovered by standard methods. The system with iron as collector requires temperatures 200 °C higher, and the iron must be slagged to enrich the noble metals by oxidation. The remaining iron is removed by treatment with hydrochloric or sulfuric acid after grinding. The advantage of iron over copper as a collector lies in the much better distribution of noble metals between metal and slag and thus in the higher degree of recovery [241].

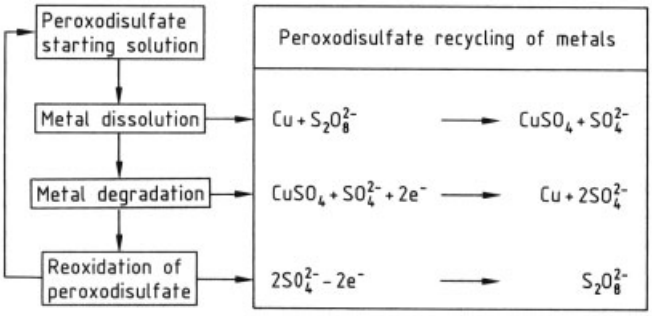

Figure 74. Block diagram of metal reclamation with peroxodisulfate for the example of copper [243]

Another route for circuit board processing has been proposed by Norddeutsche Affinerie. Mechanical and thermal treatment is omitted; the boards are fed directly to the copper converter. Metals such as copper, nickel, and the noble metals are collected in the copper matrix and refined by conventional means. Organics burn off, and ceramic constituents report to a slag. Conditions in the converter and the downstream off-gas cleaning system are such that the sum of the toxicity equivalents of dioxins is < 0.1 ng/cm^3 [240].

7.5.4.2. Material-Specific Recycling

Material-specific recycling (Fig. 69, box 8) is characterized by remelting of scrap metals to obtain well-defined materials. Metallurgical operations are not involved. Deficient alloying elements can be supplemented. The prerequisites for material recycling are that the waste materials be isolated and that their exact composition and metallic contents be known. Waste metals to be melted together must exhibit compatibility and must be free from interfering substances.

Wastes are isolated by a discrete disassembly or by continuous size reduction. The extent of disassembly or grinding must be such that the parts of the waste product can be assigned to a definite waste class. If the composition of the wastes is not known, they must be analyzed before processing. Single analyses are laborious and can only be justified for large parts; therefore, methods have been devised for automatically identifying and sorting material in small pieces, such as those produced by shredders. These techniques are based on spark spectroscopy [244], [245] and/or X-ray spectroscopy [245–247]. Once identified, the particles are removed from the belt into the proper container either pneumatically or mechanically. For 15 – 65 mm particles, identification and sorting can be performed at up to 5 t/h with a discrimination $> 96\,\%$. Sorting by chemical composition may be preceded by sorting based on physical properties (magnetic, eddy- current, dense-media, or other separation methods) so that the resulting substance classes, such as ferrous metals, aluminum and aluminum alloys, copper and copper alloys, can be further sorted on the basis of chemical composition. Similar results can be achieved by replacing the combination of physical and chemical sorting by a multistage chemical method.

A waste class is characterized by the fact that a construction material can be obtained from it by melting without addition of primary metal. The materials belonging to such a class are said to be compatible, meaning that none of them has contents of alloying elements and impurities that exceed the maximum value for the construction material to be produced by melting. For example, the waste class AlMg includes the materials Al99.9Mg1, AlMg1, AlMg2.5, and AlMg3. Compatibility is also given if a construction material can be obtained from the wastes of several waste classes by melting without addition of primary metal. This is possible when alloying element or impurity contents exceeding the maximum values in one waste class are offset by lower contents in another waste class. The technique of "charge calculation" makes it possible to determine the optimal proportions of given wastes for producing a specified material of construction by melting to be calculated [248]. The question of compatibility also

arises with surface- coated wastes. A distinction must be made between coatings that dissolve in the melt, those that enter the slag, and those that are transported into the gas phase. The distribution of coating material between melt, slag, and gas phase depends on temperature and pressure, oxygen potential, and charging method.

7.5.4.2.1. Steel

World output of steel in 1990 was 770×10^6 t; 15 % was made in Siemens – Martin furnaces, 28 % in electric-arc furnaces, and 57 % in converting mills [249].

On average, making 1 t of steel required some 400 kg of scrap, 720 kg of pig iron, and 12 kg of iron sponge. Of the scrap used, 50 % goes to the electric-arc furnace, 20 % to Siemens – Martin furnaces, and 30 % to oxygen converters [250]. Scrap reprocessing began when Martin used plant scrap in 1865. The adoption of continuous casting, which accounted for 92 % of steel produced in 1991, resulted in a drop in plant scrap from 250 to 105 kg/t of crude steel in the output in Germany between 1960 and 1991 [251]. Total scrap iron production in Germany was ca. 20×10^6 t in 1989; 11 % of this was plant scrap in foundries, and 34 % was plant scrap steel mills and rolling mills [252]. In the same period, ca. 1.1×10^6 t of scrap iron was discarded with household waste, of which 564 000 t was recycled. This reclaimed scrap contained 275 000 t of tinplate scrap, over 50 % of which was recycled for tinplate packagings [253]. In all, 12×10^6 t of scrap, including 4×10^6 t of plant scrap was used in the production of 40×10^6 t of crude steel [254]. The scrap is used as a coolant in the oxygen converter and for the same purpose in ladle metallurgy. Simple scrap remelting is carried out in foundries. Up to 30 % scrap is used in converter steel mills; in Germany, 177 kg of scrap is used in making 1 t of crude steel (102 kg of plant scrap and 75 kg of external scrap) [251].

The electric-arc furnace can handle up to 100 % scrap. The Siemens – Martin process is no longer practiced in Germany. The recycling rate for ferrous metals in 1990 was 32 % in Western Europe and 70 % in Japan. If the fraction in Western Europe were also brought up to 70 %, steel output by the classical blast furnace/converter process would fall from 95×10^6 t to 52×10^6 t, while the share of the electric-arc furnace would rise from 45×10^6 to 88×10^6 t [249].

More than two-thirds of external scrap in Germany is inlet to the arc furnace, accounting for 17 % of steel output [251]. The electric steelmaking plant generates 13.3 kg of dust per tonne of crude steel. If this dust contains > 24 % zinc, it can be refined in nonferrous metal smelters. This is not a problem for electric furnaces, but the basic oxygen furnace puts out dusts with only 2 – 3 % zinc and 0.3 % lead [254]. In 1989, the total amount of strip and sheet coated in Germany was 3.8×10^6 t; 77.1 % received metallic coatings (57.1 % galvanized, 17.3 % tinplated). With regard to melting of the resulting wastes, it was assumed that lead and zinc vaporize and thus dissolve in the steel, aluminum reports to the slag, and virtually all copper and tin present are dissolved in the steel (Section 7.5.4.1.1). Tests yielded the following average values for the percentages of these elements dissolved in the steel: for the arc furnace, 88 % of copper and 58 % of tin; for the converter, 92 % of copper and 58 % of tin; in the crucible-type induction furnace, 24 % of lead, 19 % of zinc, 97 % of copper, and 93 % of tin. Lead and zinc levels could not be determined in either the arc furnace or the converter, but it can be assumed that the contents of lead and zinc in the steel will be much lower than in the crucible induction furnace because of the high temperatures [254]. Copper and tin levels in the steel must be kept low because even low levels of copper allow the steel to become brittle, especially when it is subjected to hot stress in forming [255]. Detinning of tinplate scrap may be necessary, particularly if the target recycling rate of 70 % is reached.

7.5.4.2.2. Aluminum

German aluminum output in 1988 was 1 275 000 t, of which 744 000 t was from primary production and 530 700 t from secondary production. Total consumption in the same period was 1 672 000 t of aluminum [256]. Figure 75 shows the principal mass streams and processes used to make 1000 kg of secondary aluminum in the Federal Republic of Germany in 1988. Steps upstream of the salt-bath furnace are furnace roasting or melting (depending on

the feed material), mechanical size reduction, physical sorting, and compression if appropriate. In Germany, aluminum scrap is melted almost exclusively in rotary drum furnaces under a salt cover consisting of ca. 70 % NaCl and 30 % KCl with addition of CaF_2. The salt slag protects the aluminum against oxidation and absorbs impurities as well as decomposition and oxidation products. This process generates 0.5 – 0.7 t salt slag per tonne of secondary aluminum; the salt slag is reprocessed (see Section 7.5.5). Cast alloys are produced from wrought alloys with loss of magnesium. The amount of recycled scrap in Germany, however, has already surpassed cast alloy production, so that wrought alloys are increasingly being made into new wrought alloys. These wrought alloys contain magnesium, manganese, and silicon as principal alloy constituents; some additionally have zinc or copper. The sum of all alloying elements is around 2 – 2.5 %. The main alloying elements in cast alloys are silicon, copper, and magnesium, making up some 12 % in all [248].

Some aluminum scrap is painted or has other organic coatings. This scrap cannot be fed directly to the rotary drum furnace; the coatings must be removed by roasting. The problem is that this process takes place very slowly at low temperatures and yields inconsistent results. At high temperatures, however, there is the risk of oxidizing the aluminum. In the United Kingdom, a fluidized-bed rotary furnace has been developed for removing paint, plastics, and other combustible materials from aluminum. The heated fluidized bed consists of an inert material such as Al_2O_3. The gas produced is burned in a final combustion chamber. Gränges in Sweden employs a two- chamber furnace in which the organic components in the scrap are roasted in the first chamber, and the roasting gases are burned in the second. The energy consumption for roasting and melting is partly covered by the roasting gases [257]. On the basis of a U.S. patent, Vereinigte Aluminiumwerke (VAW) has devised a three- chamber furnace for processing painted, coated, and oil- contaminated materials. Melting capacity is 3 t/h; the furnace holds 90 t, of which 70 t is left in the unit to ensure that the passages between chambers remain below the bath surface so as to separate the different gas atmospheres in the chambers. This furnace is run as a system isolated from the outside atmosphere. The energy content of the roasting gases is used for the system, and the heat content of the flue gases is recovered and utilized to preheat the combustion air. Oxidation of the aluminum is prevented by not allowing the scrap to come in contact with any flame. Energy is supplied in the main hearth, not the melting hearth [253].

7.5.4.3. Recycling of Material Wastes

Metal waste (scrap waste) that is not subject to material-specific recycling is inlet to the third variant of product waste recycling, namely recycling of material wastes (Fig. 69, box 10). Part of this stream may originate in product-specific recycling. The cost per unit of metal reclaimed is higher in this variant than in product-specific or material-specific recycling. Recycling of material wastes is characterized by metallurgical removal and separation of substances.

Aluminum is a relatively base metal, and so only a few elements can be removed from molten aluminum by refining. Aluminum scrap is therefore selected to have compatibility for material-specific recycling, so that the desired construction material can be obtained without refining, apart from the removal of a few impurities. In contrast to aluminum, copper is a relatively noble metal and can readily be separated metallurgically from other elements. Copper and copper alloy scrap from which construction materials can not be made by direct remelting is forwarded to recycling of material wastes.

World copper consumption in the year 1991 was 12.18×10^6 t. Of this, 38.3 % was produced from secondary raw materials. In Germany, 1989 consumption was 1.128×10^6 t of copper, 553 000 t (49 %) of which was made from copper- containing scrap [221], [233]. Copper wastes can consist of copper compounds (mostly oxides) or be metallic. Metallic copper wastes are classified as copper scrap generally containing more than 96 % copper, copper alloys (bronze, red and yellow brass, nickel silver, etc.), and copper composites (radiators, motors, armatures, etc.). These wastes are subjected to a variety of metallurgical operations, depending on their compositions.

Figure 76 presents the flowsheet of Hüttenwerke Kaiser [222], [233]. Most of the copper is reclaimed by pyrometallurgy. The

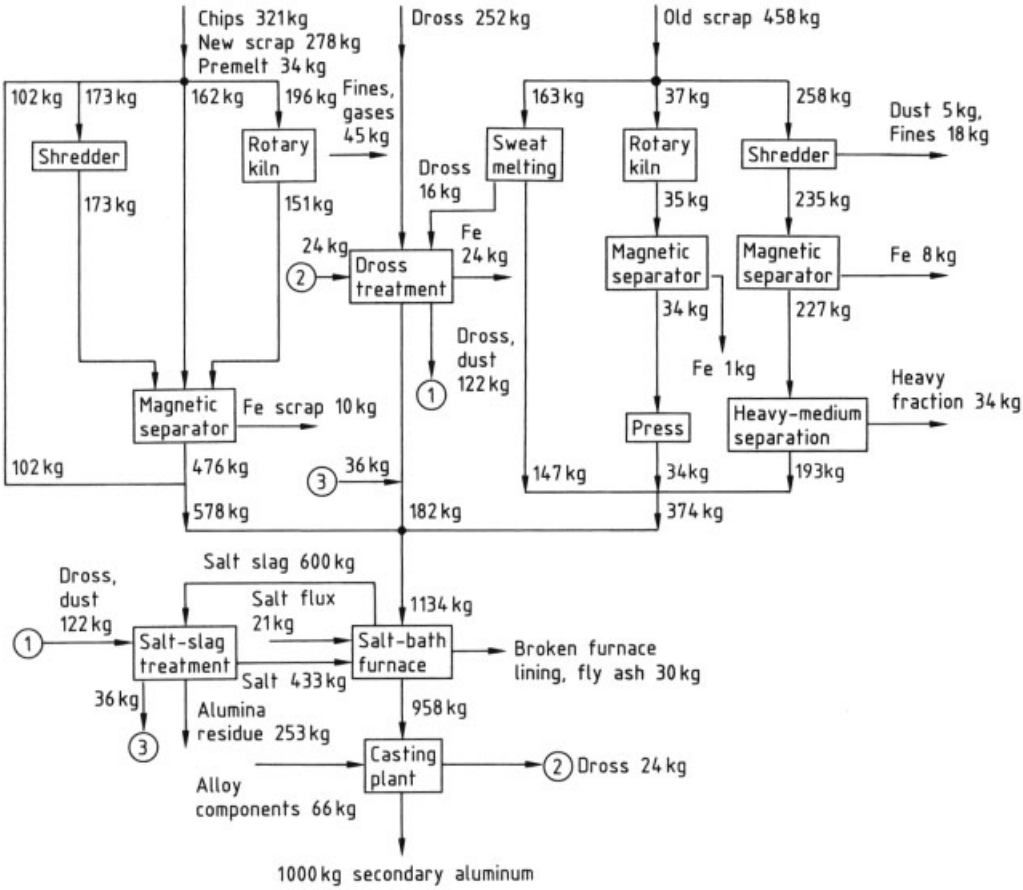

Figure 75. Material streams for production of 1000 kg of secondary aluminum [256]

main processes are reduction and refining. Operations such as cementation and reductive electrolysis are not important in the recycling of copper. Shaft furnaces, electric furnaces, and top-blown rotary converters (TBRC) are employed for reduction; both stationary and tiltable reverberatory hearth furnaces as well as rotary drum furnaces are used in refining. Feeds to shaft furnaces range from low-copper dusts, ashes, and other fine-grained materials and nonmetallic residues to shredder material and copper – iron materials, and also include various slags and residues from other copper-enrichment processes. Shaft furnace feeds are divided into tin-rich and zinc-rich materials. The shaft furnace, like the blast furnace, operates in a countercurrent fashion: Alternate portions of coke and burden are supplied from the top while air is injected through tuyeres at the bottom to effect combustion of the coke and generate the proper furnace atmosphere. In contrast to the blast furnace, the shaft furnace has a "fire top" to oxidize vapors of tin, lead, and especially zinc in the furnace. These oxides then pass to the off-gas filtration system. The metal phase forms in the lower furnace; depending on the burden, it contains ca. 70 % copper along with tin, lead, zinc, nickel, and iron. A $CaO - SiO_2 - FeO$ slag is also pro-duced. Metal and slag can be tapped at around 1300 °C, either continuously or intermittently. Metals produced in the shaft furnace with zinc-rich burden contain over 80 % copper; these, unlike tin- and lead-rich metal, can be forwarded directly to the anode furnace for preliminary refining. Both kinds of shaft fur-

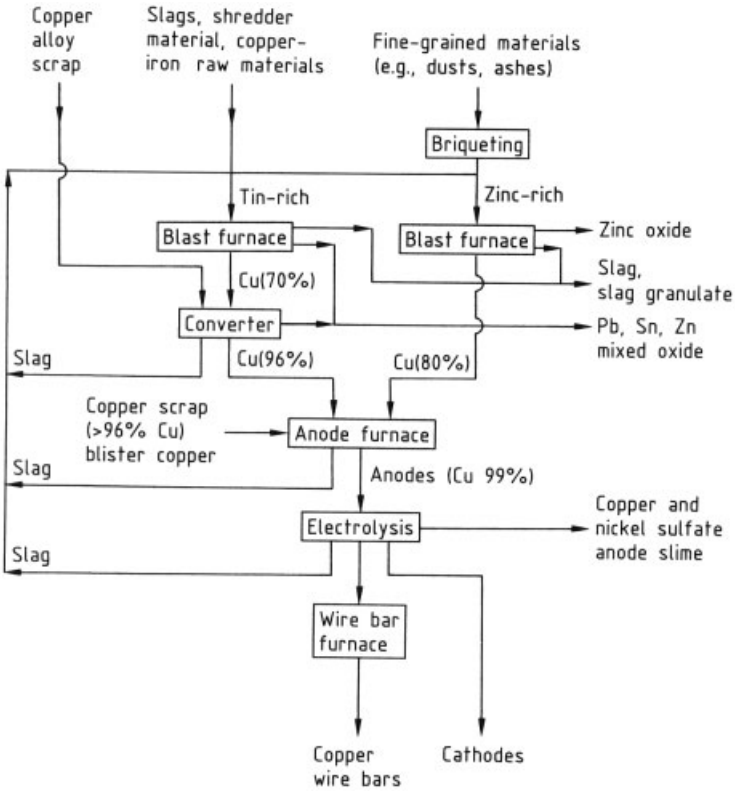

Figure 76. Block diagram of a secondary copper smelter [222]

nace produce an inert slag, which is salable as a building material.

The top-blown rotary converter can be used for the same operation as the shaft furnace. The TBRC always operates in batch mode. The charge is melted by a torch fed with gas, fuel oil, and also powdered fuels. The metallurgical process is similar to that in the shaft furnace, but mass transport is much more intensive because of the rotation and the blowing of combustion gases onto the melt. The fuel/air ratio in the TBRC can be used to set up a reducing atmosphere, as in the shaft furnace, or an oxidizing atmosphere, as in the converter. This equipment is thus more flexible than the shaft furnace [222].

The original Knudsen patent application on the conventional scrap-metal converter for refining copper alloy scrap was filed as early as 1915. In the converter, a horizontal cylinder with tuyeres on the sides, coke is burned to melt the copper alloy scrap (yellow and red brass, bronze, etc.) together with tin- and lead-rich copper from the shaft furnace. When air is blown through the melt, some of the inlet tin, lead, and zinc is volatilized, and is recovered as flue dust in the filter. At the end of the charge, a metal containing ca. 96 % copper, some of the inlet nickel, and the noble metals is left in the converter. This metal is forwarded to the anode furnace for rerefining. In addition to iron, the slag contains part of the inlet valuable metals and is therefore returned to the shaft furnace. The mixed tin-, lead-, and zinc- containing oxides from the shaft kiln and converter are fed to the reverberatory furnace or rotary reverberatory furnace, where they are melted under reducing conditions with soda and coal fines; the zinc is volatilized and is then oxidized to zinc oxide. The lead – tin alloy is used as a starting product for soldering tin, type metal, and other metals. The zinc oxide can be sold as a paint pigment. The zinc oxide from the shaft kiln is less pure and is therefore delivered to a zinc smelter for zinc recovery [258], [259].

Copper from the converter and the shaft kiln for zinc-rich material are melted together with copper scrap containing over 96 % copper in the anode furnace (stationary or rotary reverberatory furnace) and refined. Refining is effected by blowing in air; the next step is reduction by poling with tree trunks or injecting natural gas. Of metallic impurities, aluminum, zinc, and iron can be decreased to < 0.01 %, lead and tin to ca. 0.2 %. Nickel and noble metals, in contrast, virtually cannot be removed by pyrometallurgy. The metal produced in this way, with about 99 % copper, is cast into anodes for electrolytic refining.

In this electrolysis, the anodes are dissolved in sulfate electrolytes when a d.c. voltage is applied. The baser elements, such as nickel, arsenic, zinc, and iron, go into solution in ionic form along with copper or are, like lead, precipitated as sparingly soluble compounds. Elements more noble than copper are not dissolved. Noble metals such as silver, gold, and platinum group metals accumulate in the anode slime with the sparingly soluble compounds. The noble metals, present at very low concentrations in the feed, are enriched to 7.5 % in the anode slime [233]. Apart from the noble metals, tin, antimony, arsenic, selenium, and tellurium are recovered from the anode slime. In principle, only copper is deposited on the cathode in a purity of at least 99.95 %. The secondary copper obtained in this way is of a quality equivalent to copper derived from primary resources. The impurities accumulating in the electrolytes are removed by electrolyte conditioning. In producing 116 000 t of cathodic copper from secondary raw materials, Hüttenwerke Kaiser obtained 2300 t/a of nickel sulfate and 2800 t/a of copper sulfate by fractional crystallization and electrolytic copper recovery from electrolytes. The sulfate removed in these operations must be made up by adding 78 % sulfuric acid to the electrolytes [222], [233].

The cathodes are then melted in the wire bar furnace and the copper is either cast into wire bars or continuously cast into long formats.

7.5.5. Recycling of Residues

In residue recycling (Fig. 69, box 12), nonmetallic wastes obtained in various other forms of recycling are used to produce metals. Residues from production are treated by production waste recycling (box 2). The methods used to treat production residues are similar to those for recycling residues, but the two types of residues usually differ in that recycling residues are more contaminated and/or intermingled. The effort and costs for obtaining metals of a given purity are greater for recycling of recycling residues than for recycling of production wastes. Residue recycling deals only with residues from recycling.

Electric Arc Furnace Fly Ash. About one-third of the zinc consumed in the Federal Republic of Germany in 1987 was used for surface coatings [260]; this corresponds to 2.8×10^6 t of steel (strip and sheet) receiving zinc coatings. In the United States and Western Europe, the processing of galvanized sheet steel in the electric arc furnace leads to zinc- and lead-containing dusts, each in a quantity of 500 000 t/a [220]. Of this dust, ca. 880 000 t, containing $> 150\,000$ t of zinc was recycled in the Western world [261].

The rotary-furnace volatilizing process (Waelz kiln) was established for processing electric steelmaking plant dusts with 10 – 35 % zinc and 3 – 6 % lead. This process features continuous operation in an inclined rotary furnace, the use of solid reducing agents, and the avoidance of liquefaction of the furnace contents. Zinc-containing feed materials are reduced by the added carbonaceous materials; the resulting zinc vapor is reoxidized in the furnace and carried out with the flue gas as ZnO. Lead is volatilized in the metal form and also in the sulfide form, the oxide, or the chloride. The product oxide, containing 54 – 58 % zinc and 8 – 12 % lead, is forwarded to zinc smelting. The slag, with an FeO content of 34 – 38 %, can be used for roadbuilding and as construction material. Before being delivered to the Imperial Smelting (IS) process or the lead – zinc blast furnace, the oxides produced are agglomerated. Hot briqueting (Fig. 77) has been used for this purpose by Berzelius since 1975. Virtually all the oxides in the feed material can be briqueted at ca. 500 °C without added binders [219]. The briquetts are charged together with lumpy sinter of primary raw materials at 200 °C, coke preheated to 800 °C, and small amounts of lime as a slag-forming agent. The air needed to com-

bust the reducing agent is preheated to 1150 °C. After reduction in the shaft, lead bullion and slag are tapped at the hearth of the furnace. The top gas discharged from the furnace at about 1000 °C contains ca. 6 – 8 vol % zinc vapor, 10 – 12 vol % CO_2, and ca. 20 vol % CO. The zinc vapor is condensed to liquid in a lead-spray condenser at the head of the furnace and is withdrawn continuously. The zinc dissolved in the hot lead in the condenser is separated from the lead melt in the cooling trough and at lower temperature in the softening furnace; the lead is returned to the condenser. The resulting crude zinc contains 98 % Zn [262]. In addition to flue dusts from electric steelmaking plants, a wide variety of other zinc- containing residues can be handled in the IS blast furnace (Fig. 77).

Salt Slags. Virtually all the 530 000 t of secondary aluminum produced in Germany in 1988 came from rotary drum furnaces. For each tonne of secondary aluminum, between 0.5 and 0.7 t of salt slag was also produced, depending on the degree of contamination of the feed aluminum scrap. The resulting output of salt slag was ca. 270 000 t. In addition some 75 000 t of widely varying drosses and dusts occurred in ball mills, generated by conditioning of aluminum drosses [263]. About 600 000 t of salt slag was produced in Western Europe in 1990 [220].

Today this salt slag is treated by the B.U.S.-Chemie solution- crystallization process. The steps in this process are crushing, screening, and dry separation of aluminum metal; dissolving, degassing, and preparation of a solid-liquid suspension; separation of off-gases, thickening, filtration of insoluble matter, recovery of alumina; and crystallization, centrifugation, and recovery of recycling salt. On the basis of an average salt/slag ratio of 600 kg per tonne of secondary aluminum, this means that on average the conditioning of salt slag corresponding to 1000 kg of secondary aluminum yields 42 kg of aluminum alloy pellets, which can be directly forwarded to remelting plants; 167 kg of recycling salt, which can be used as cover salt in rotary drum furnaces; and 196 kg of alumina residue containing heavy metals in a nearly insoluble form, which can be disposed of in landfills. The gases evolved on dissolution (relative to 1000 kg of secondary aluminum produced) contain 4.5 cm^3 H_2, 2.0 cm^3 CH_4, 0.31 cm^3 NH_3, 0.02 cm^3 PH_3, and 0.0001 cm^3 H_2S. The gases are treated with activated carbon; PH_3 and H_2S are adsorbed on the carbon, contacted with diluting air, and thus oxidized so that the treated gas is in compliance with emission limits [263], [264].

7.6. Sludges

7.6.1. General

Sludges vary widely in character and are treated or disposed of for a variety of purposes. As a consequence, sludge treatment imposes stringent requirements on the physical, chemical, and biological processes involved: transformation of matter, solid – liquid separation, and thermal treatment. Associated with these basic operations are the secondary problems of controlling liquid and gaseous emissions.

The bulk of sludges comprises sewage sludges, industrial production sludges, water treatment sludges, dredging sludges, and mining sludges.

Because treatment of sewage sludge is by far the most serious problem today, the discussion that follows is concerned with this topic; the other types of sludge and problems relevant to them are mentioned only in passing.

Most of the techniques used for sewage sludge treatment can be employed in identical or modified form for other sludges.

The treatment of slurries and sludges in production processes must be optimized with regard to process and product quality, and most known process sequences go through a long development period before they are perfected. When sludges are disposed of, legislation and regulations governing the returning of sludge components to the environment must be borne in mind. Legislative and regulatory provisions can lead to technical guidelines that bar the use of proven techniques and mandate conversion to new processes.

7.6.2. Quantities of Sewage Sludge Treated

The amount of sludge produced depends on the effectiveness of wastewater treatment. A definite quantity of sludge can be anticipated for a given treatment method.

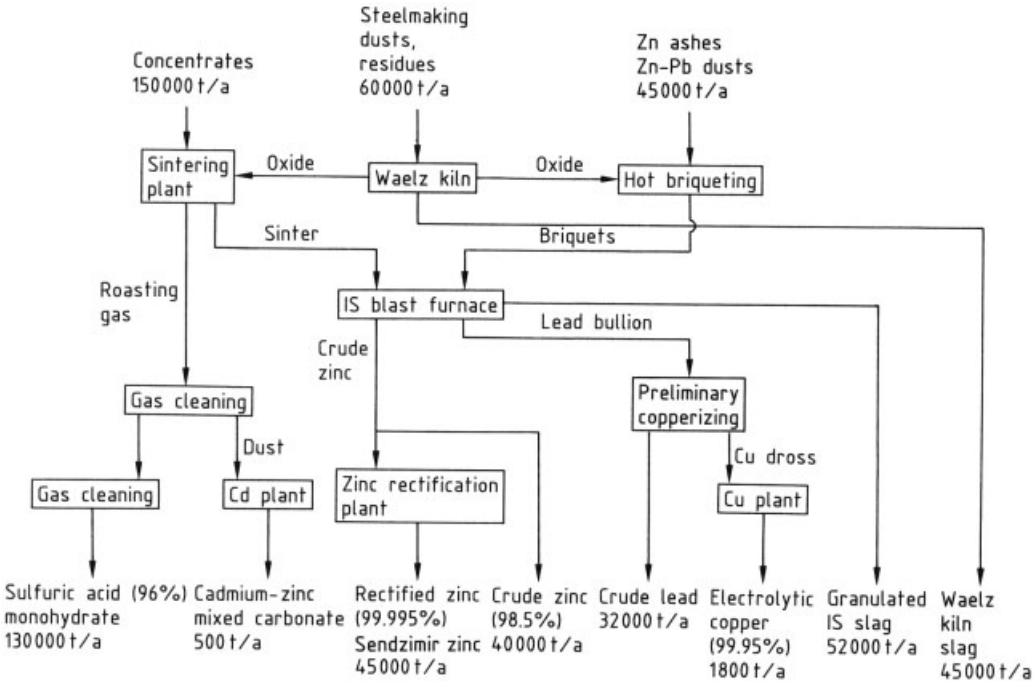

Figure 77. Processing of steel-mill dusts and other Zn- and Pb- containing secondary materials together with primary raw materials [219], [262]

Advanced wastewater treatment practices, in particular nitrification/denitrification and precipitation of phosphates, are expected to result in appreciable changes in both the specific quality and the volume and mass flow rates of sewage sludges.

The precise extent of these changes in per capita sludge solids and sludge volume cannot be statistically calculated today because not enough representative data have been collected.

It is, however, clear that the increase in amount of solids and the increase in volume are not synchronized. In phosphate precipitation especially, the choice of process and of precipitating agent have a substantial effect on the per capita quantities.

With regard to nitrification/denitrification, no significant increase in sludge solids is anticipated; there may in fact be a reduction of up to 10 % in the per capita quantity of sludge solids. The per capita liquid sludge volume is likely to rise by 10 – 20 %, however.

With the simultaneous precipitation of phosphates as now favored, a 20 % increase in the per capita amount of sludge solids can be expected.

The rise in per capita volume depends on the type of biological treatment process and the precipitating agent used; on the whole, an average rise of ca. 10 % range can be anticipated. In general, it can be assumed that the use of upstream or downstream precipitation tanks instead of simultaneous precipitation will result in markedly greater increases in the per capita solids quantity and sludge volume.

If the two advanced wastewater treatment processes are combined, the percentage rise in per capita sludge volume will be greater than the percentage increase in sludge solids per person.

A rough figure for the quantity of sewage sludge generated per person (or person-equivalent) per day is 2 L, with a solids content of ca. 4 %.

Calculations of sludge quantities have to take account of project-specific features having to do with the sewer system type (combined or separate) and the character and size of industrial users.

At present, the quantity of sewage sludge to be treated in Germany is ca. $(65 - 78) \times 10^6$ m^3/a,

with an average solids content in the range of 4.2 – 4.6 %.

7.6.3. Requirements for Sludge Recycling and Disposal

The ultimate goal of sewage sludge treatment is reclamation or disposal of residues contained in the sludge.

The dominant disposal options at present are agricultural use and landfill disposal after mechanical dewatering.

German legislation sets some rules for these general reclamation/disposal approaches.

The modalities of the Sewage Sludge Regulation (Klärschlammverordnung) [265] govern the agricultural use of sewage sludge.

The most important limits set by this regulation apply to the content of heavy metals in sludges approved for agricultural use and in the soils to which they are applied (see Table 14).

Table 14. Heavy metal limits in sewage sludges and soils when sewage sludge is used agriculturally in accordance with German sewage sludge regulations

Metal	Concentration in sewage sludge, mg/kg∗	Concentration in soil, mg/kg soil
Lead	900	100
Cadmium	10/5∗	1.5/1∗
Chromium	900	100
Copper	800	60
Nickel	200	50
Mercury	8	1
Zinc	2500/2000∗	200/150∗

∗ Dry matter. ∗∗ For application to light soils with < 5 % clay and soils with pH between 5 and 6.

The deposition of sewage sludges in landfills of various categories is subject to the criteria of the Technical Directive on Municipal Waste [266] (see Table 15).

The limitation on organic content in dry sludge, determined as loss on ignition, means in practice that sludges cannot be placed in landfills without incineration as a treatment step.

In a transition period extending until the year 2005, the method of sewage sludge disposal that is by far most commonly practiced now – landfill disposal after mechanical dewatering – will continue to be permitted. Local landfill operators, however, are requiring improved strength values for sludges to be disposed of (see Table 15).

7.6.4. Sewage Sludge Treatment Processes

The purpose of sewage sludge treatment is to prepare raw sludges differing in origin and composition for reuse or ultimate disposal. A number of processes and combinations of processes have been developed.

The most important combinations now used in industrialized countries are summarized in Figure 78 [267].

Once the quality requirements of the Technical Directive on Municipal Waste come into full force, the combinations numbered 6, 7, 13, 14, 19, and 20 will be removed from the list.

7.6.5. Unit Operations in Sewage Sludge Treatment

7.6.5.1. Thickening and Dewatering

Sewage sludges have water contents of ca. 90 – 99.5 % depending on origin. The high water contents and the resulting large sludge volumes cause problems in some sludge treatment processes. In addition, many sludges have a high water-retention capacity, arising from various kinds of binding force, which makes the removal of water difficult.

The most important unit operations in sewage sludge treatment thus include thickening and dewatering.

Sludge thickening can be performed with the aid of gravity or by centrifugation. This removes chiefly interstitial water, which is retained only by weak forces. Thickening of sewage sludges can also be effected by flotation or with filters.

Higher pressure differences (positive pressure or vacuum) in mechanical dewatering or more intensive centrifugation can overcome the stronger forces binding contact and capillary moisture to the solids. Figure 79 shows how the technologically and economically optimum volume reduction can be achieved by removal of sludge water in these types of equipment and in the thermal processes of drying and incineration.

The thickenability and maximum dewaterability of sewage sludge are listed in Table 16. Table 17 presents attainable results in

Table 15. Landfill classification criteria according to the Technical Directive on Municipal Waste, Annex B

Number	Parameter	Criterion values	
		Landfill class I	Landfill class II
1	Consistency [a]		
1.01	Vane shear strength	$\geq$ 25 kN/m^2	$\geq$ 25 kN/m^2
1.02	Axial deformation	$\leq$ 20 %	$\leq$ 20 %
1.03	Uniaxial compressive strength	$\geq$ 50 kN/m^2	$\geq$ 50 kN/m^2
2	Organic part of dry residue of original substance [b]		
2.01	As loss on ignition	$\leq$ 3 wt %	$\leq$ 5 wt %
2.02	As TOC	$\leq$ 1 wt %	$\leq$ 3 wt %
3	Extractable lipophilic substances in original substance	$\leq$ 0.4 wt %	$\leq$ 0.8 wt %
4	Eluate criteria		
4.01	pH	5.5 – 12.0	5.5 – 12.0
4.02	Conductivity	$\leq$ 6 000 µS/cm	$\leq$ 50 000 µS/cm
4.03	TOC	$\leq$ 20 mg/L	$\leq$ 100 mg/L
4.04	Phenols	$\leq$ 0.1 mg/L	$\leq$ 50 mg/L
4.05	Arsenic	$\leq$ 0.1 mg/L	$\leq$ 0.5 mg/L
4.06	Lead	$\leq$ 0.2 mg/L	$\leq$ 1 mg/L
4.07	Cadmium	$\leq$ 0.05 mg/L	$\leq$ 0.1 mg/L
4.08	Chromium(VI)	$\leq$ 0.05 mg/L	$\leq$ 0.1 mg/L
4.09	Copper	$\leq$ 1 mg/L	$\leq$ 5 mg/L
4.10	Nickel	$\leq$ 0.2 mg/L	$\leq$ 1 mg/L
4.11	Mercury	$\leq$ 0.005 mg/L	$\leq$ 0.02 mg/L
4.12	Zinc	$\leq$ 2 mg/L	$\leq$ 5 mg/L
4.13	Fluoride	$\leq$ 5 mg/L	$\leq$ 25 mg/L
4.14	Ammonium-nitrogen	$\leq$ 4 mg/L	$\leq$ 200 mg/L
4.15	Chloride	$\leq$ 500 mg/L	$\leq$ 5 000 mg/L
4.16	Cyanides, easily released	$\leq$ 0.1 mg/L	$\leq$ 0.5 mg/L
4.17	Sulfate [c]	$\leq$ 500 mg/L	$\leq$ 1 400 mg/L
4.18	Nitrite	$\leq$ 3 mg/L	$\leq$ 6 mg/L
4.19	AOX	$\leq$ 0.3 mg/L	$\leq$ 1.5 mg/L
4.20	Water-soluble fraction (residue on evaporation)	$\leq$ 3 wt %	$\leq$ 6 wt %

[a] 1.02 together with 1.03 can be used as equivalent to 1.01. The consistency must be determined for each case in accordance with statistical requirements for landfill stability. 1.02 combined with 1.03 represents a lower boundary especially for cohesive, fine-grained wastes.
[b] 2.01 can be used as equivalent to 2.02. Requirement does not apply to contaminated soil that has been excavated and is to be placed in a hazardous-waste landfill.
[c] Limit may be exceeded up to 1 400 mg/L SO_4^{2-} for gypsum rubble and gypsum construction waste as well as other gypsum-containing building rubble, provided the Ca^{2+} concentration in the eluate is at least 0.43 times the SO_4^{2-} concentration.

Table 16. Thickenability and dewaterability of sewage sludge (in % solids content) [258]

Water-retaining properties of sludge	Thickenability[a]	Dewaterability		
		Continuous belt filter,[b] and centrifuges[c] polymer conditioning	Filter-presses, metal salt or polymer conditioning	
			with lime/additives	without lime/additives
Good thickenability/dewaterability	>7	>30	>38	>45
Moderate thickenability/dewaterability	4 – 7	18 – 30	28 – 38	35 – 45
Poor thickenability/dewaterability	<4	<22	<28	30 – 35[c]

[a] Static thickening without conditioning. [b] For initial solids content 3 – 9 %. [c] Higher values can be obtained with high-performance centrifuges.

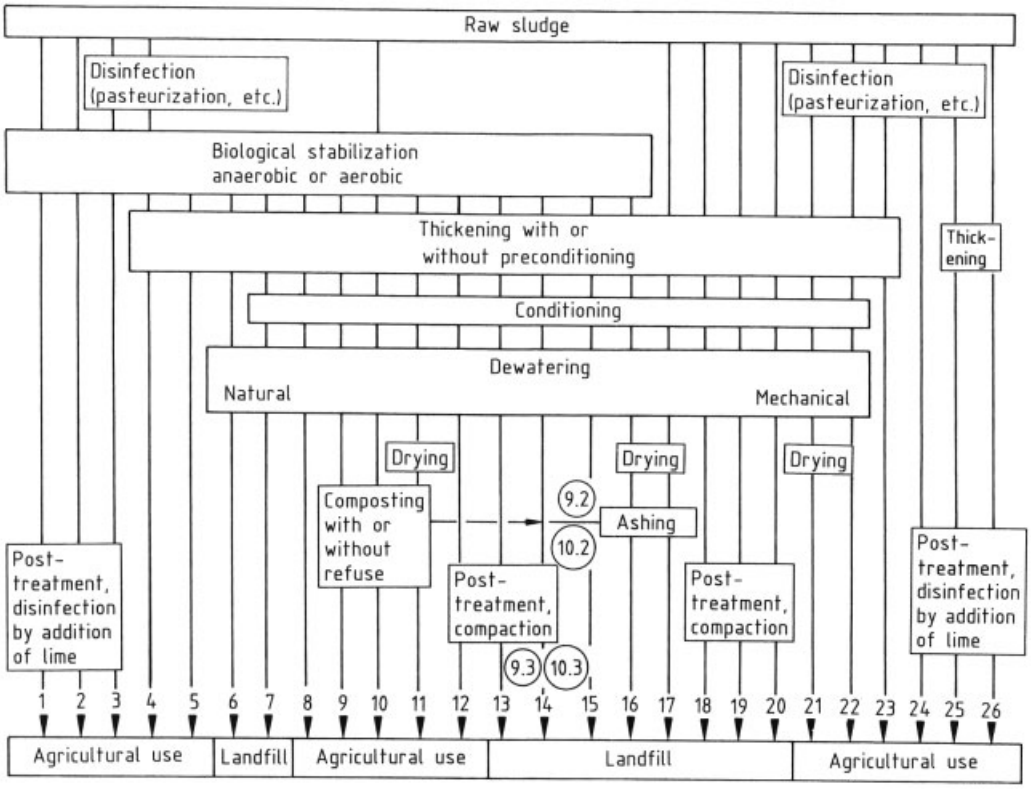

Figure 78. Possible combinations of processes for treatment, recycling, and disposal of sewage sludges

mechanical dewatering of sewage sludge when various conditioning methods are employed.

For detailed discussion of these topics, see [268].

7.6.5.2. Stabilization

DIN 4045 defines the stabilization of sludges, especially sewage sludges, as methods of sludge treatment for the purposes of:

1) Decreasing the amount of odor-forming constituents
2) Reducing the quantity of sludge solids
3) Improving dewaterability
4) Diminishing the number of pathogens

The following are distinguished:

1) Aerobically stabilized sludge
2) Anaerobically stabilized or digested sludge
3) Chemically stabilized sludge
4) Thermally stabilized sludge

The purpose of stabilizing sewage sludges is to reduce the content or modify the character of organic substances that can be attacked by putrefactive bacteria in such a way that the sludges so treated have no further appreciable noxious and polluting effects or impacts on the environment. Generally, the degradation and transformation processes associated with stabilization involve transformation of high-molecular mass sludge constituents to lower molecular mass compounds. Often the aim is mineralization, that is, their transformation to inert species. Of the possible processes for sludge stabilization, the biological methods of aerobic and anaerobic stabilization have gained the widest adoption.

One method of chemical stabilization is to add substances (chiefly lime) that raise the pH drastically, thus hindering transformation processes. This method is mainly used in the Scandinavian countries.

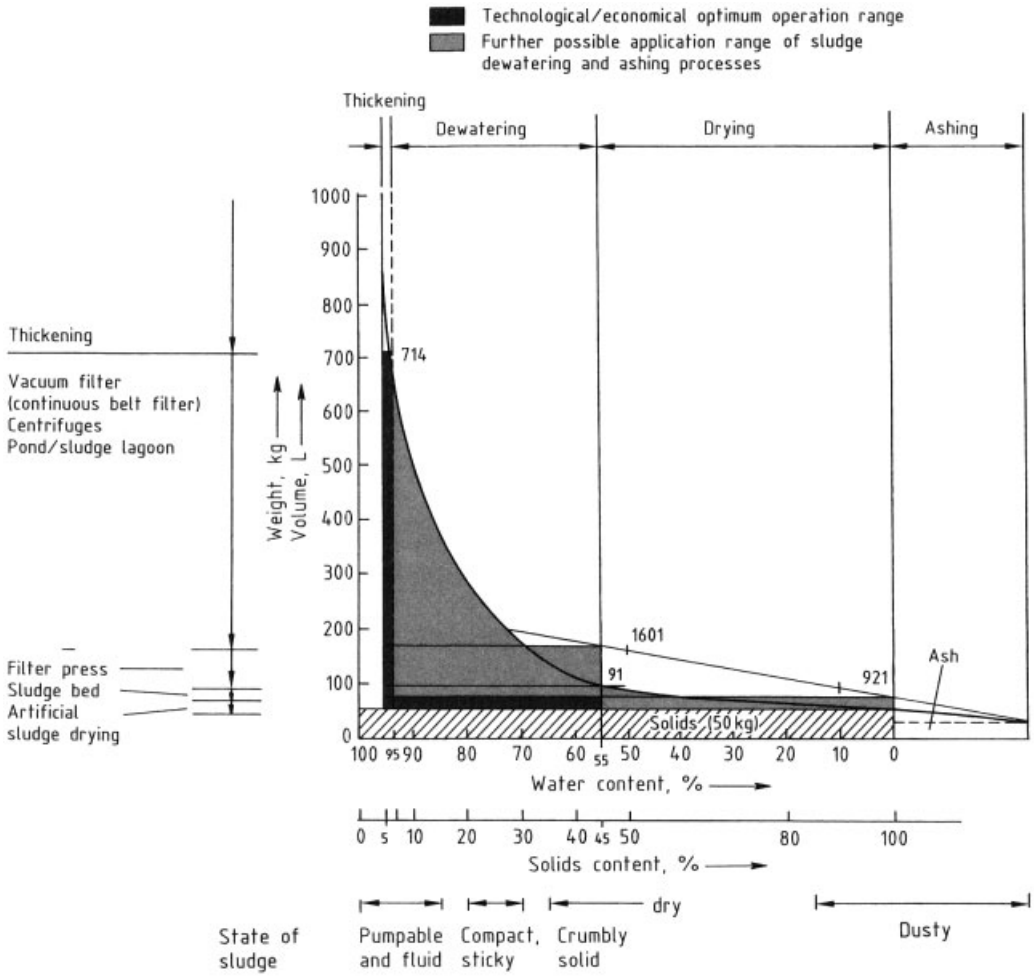

Figure 79. Technically and economically optimal volume reduction by dewatering and incineration or combustion

Formally, all thermal treatments are classified as chemical processes; these include combustion/incineration, wet oxidation, and gasification (see Sections 9.3 and 9.4).

Aerobic biological stabilization of sewage sludge has the following variants:

1) Simultaneous aerobic sludge stabilization
2) Separate aerobic sludge stabilization at ambient temperature
3) Aerobic-thermophilic sludge stabilization in the liquid phase
4) Aerobic-thermophilic sludge stabilization in the solid (or nonflowable) state of aggregation (sludge composting)

For a discussion of the principles underlying these methods, see [269].

The dominant biological process in larger treatment systems is anaerobic digestion. The microbial transformations are effected by groups of bacteria that function under air exclusion at 30 – 55 °C.

This process yields biogas, which can be used to generate energy.

Aerobic processes and digestion can be combined in so-called dual sludge stabilization. The combination of an upstream aerobic-thermophilic stage and a downstream anaerobic stage operating in the mesophilic temperature range is most widely used.

Table 17. Mechanical sludge dewatering as a function of conditioning process, under the assumption of moderate sludge dewaterability [258]

Conditioning process	Centrifuges∗ and continuous belt filter				Filter presses			
	Dry residue,∗∗ %		Laboratory transverse resistance, kN/m^2		Dry residue,∗∗ %		Laboratory transverse resistance, kN/m^2	
	Mean	Range	Mean	Range	Mean	Range	Mean	Range
Conditioning with polymers	25	18 – 30	3	1 – 5	32	28 – 38	15	8 – 30
Conditioning with metal salts	Not usual for municipal sewage sludge							
Conditioning with metal salts and slaked lime	not appropriate				40	35 – 45	25	10 – 100
Thermal conditioning	Not recommended due to secondary pollution							
Condtioning with polymers and powdered coal	Only appropriate with subsequent incineration							
Conditioning with polymers and pre- or post-treatment with quicklime, cement/fly ash	depends on disposal objective and quantity		10	3 – 30	rarely used			
Conditioning with polymers and post-treatment with ballasts	depends on disposal objective and quantity		3	1 – 5	not appropriate			

∗ Higher values can be obtained with high-performance centrifuges.
∗∗ Including conditioning agents and additives.

Figure 80 shows graphically the recommended application areas of the main biological stabilization processes with respect to plant size. Further information on dual biological stabilization can be found in [270].

7.6.5.3. Disinfection

The German Sewage Sludge Regulation no longer mandates the disinfection of sewage sludge prior to distribution on certain types of agricultural land. It is, nevertheless, a quality-enhancing practice in sludge recycling, and purchasers and users in the agricultural sector place a positive value on it.

The processes most suitable for disinfection of sewage sludge are those in which the temperature reaches the thermophilic range ($> 55\,^{\circ}$C).

These include aerobic – thermophilic stabilization in the liquid or solid state (composting). An hygienic product can also be obtained by thermophilic digestion or dual sludge stabilization.

Process conditions for proper disinfection are discussed in [271].

7.6.5.4. Conditioning

The purpose of sludge conditioning is to improve sludge thickenability or dewaterability. Sludge conditioning can be effected chemically or physically.

In chemical conditioning, coagulation and flocculation are induced or accelerated in several ways. An organic flocculant can be added to promote solid – liquid separation, or an inorganic flocculant or precipitating agent can be used to form flocs under hydrolyzing conditions. The modification or neutralization of the electrostatic charges on the sludge constituents or the creation of new chemical compounds also leads to the formation of microflocs on the solid and colloidal particles present in the sludge.

The multiplication of solid particles leads to larger agglomerates /microflocs. Because the water-retaining capacity is simultaneously reduced, sedimentation is accelerated and filterability is improved.

Physical conditioning processes fall into mechanical and thermal classes. Mechanical conditioning changes the structure of the constituents in the sludge by the addition of in-

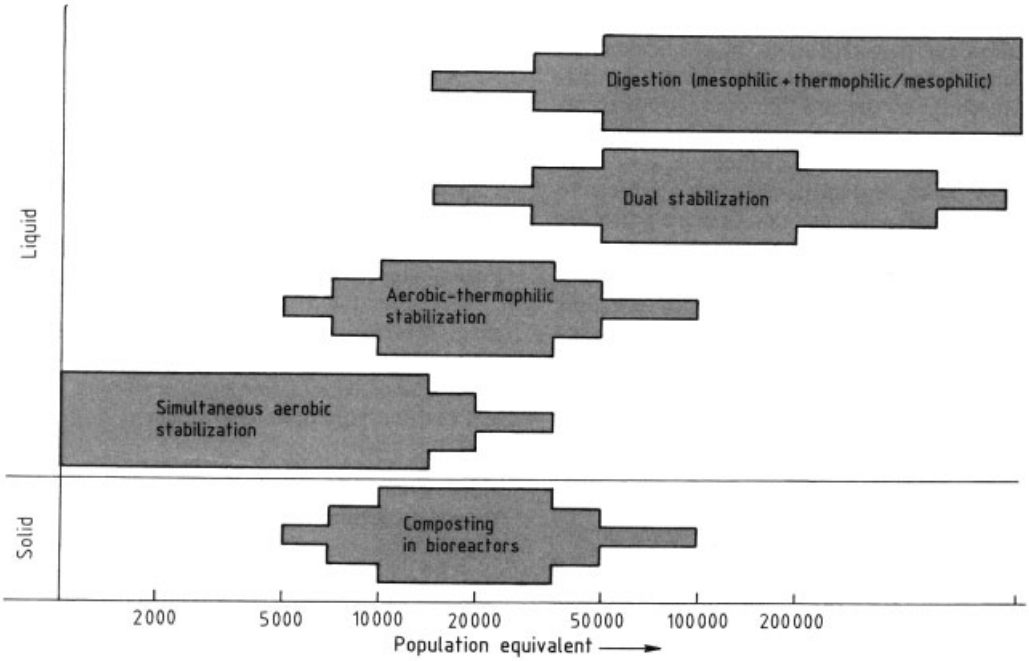

Figure 80. Recommended application ranges of principal variants of biological sewage sludge stabilization in relation to treatment plant size

organic or organic reagents that do not initiate chemical reactions. Effective additives include ash, coal dust, sawdust, sand, and infusorial earth products. In thermal conditioning, the same above mechanisms are effected by freezing or heating the sludge.

Low-temperature conditioning involves lowering the temperature to – 20 °C. This technique is used primarily in regions with a cold climate and is commonly practiced in natural (gravity) dewatering units.

Thermal conditioning methods are classified by temperature range. Low-thermal conditioning takes place between 60 and 80 °C. High-thermal conditioning is carried out between 180 and 230 °C at 10 – 25 bar. The use of thermal conditioning has declined greatly, since the process gives rise to intense odors (especially the high-temperature variant) that usually cannot be controlled.

Table 17 summarizes sludge dewatering as a function of the conditioning process.

7.6.5.5. Drying

If it is necessary to go beyond the degree of water removal given by mechanical dewatering and also remove ballast water, the only option is "thermal dewatering" or drying.

"Complete" drying makes it possible to increase the fraction of dry residue in the sludge to 90 %, sometimes over 95 %.

The extent of sludge drying and the method used will depend chiefly on the downstream sludge treatment steps or the mode of recycling or ultimate disposal. Processes fall into two major classes: convection drying and contact drying.

In convection dryers, heat is supplied by hot gases flowing through the sludge. In contact dryers, heat is delivered through heated wall surfaces. The control of emissions is a major problem in the drying of sewage sludges.

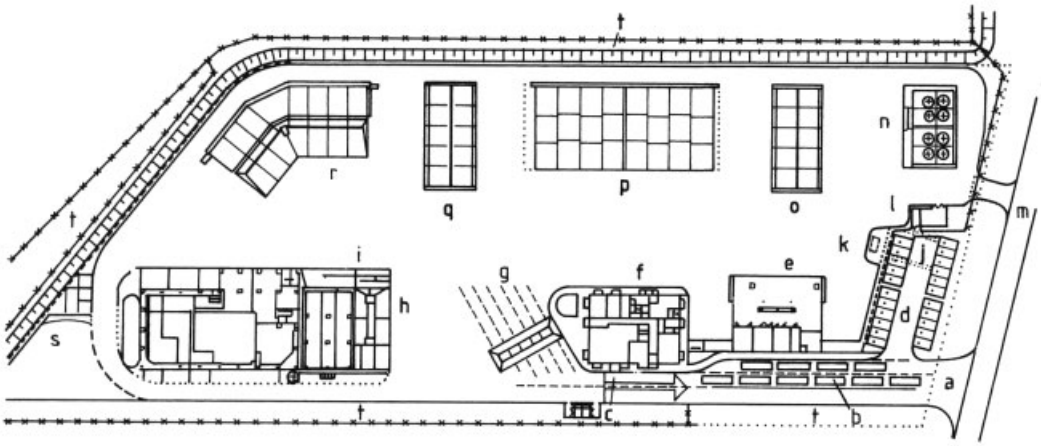

Figure 81. Intermediate storage site for hazardous wastes in Munich
a) Entrance and exit; b) Holding area for arriving trucks; c) Scales; d) Automobile parking; e) Multipurpose building; f) Operations building; g) Sampling pit; h) Car/truck wash; i) Physicochemical treatment and wastewater treatment plant; j) Rainwater and fire fighting water retention basin; k) Fueling station; l) Pump house; m) Fire department access; n) Tank farm; o) Stands for skips; p) Semitrailer stands and drum storage; q) Stands for skips; r) Solids storage pile; s) Substation; t) Greenbelt

7.6.6. Quantities of Other Sludges;Disposal Options

7.6.6.1. Industrial Production Sludges

The discussion in the preceding sections includes sludges from industrial production that are generated either in special industrial treatment plants or in municipal treatment facilities into which industrial wastewaters are introduced.

In addition, production processes generate a variety of sludges that must be disposed of. Examples are hydroxide sludges, roll-sintering sludges and metal sludges from the iron and steel industry, sludges from food and sugar production, papermaking sludges, and so on.

The potential quantities of these can be determined only approximately, because in many cases the disposal of such sludges has never been regarded as a problem. The total quantity of these production sludges in Germany is estimated at $(40-50)\times10^6$ m^3/a. This quantity is associated with a dry residue of about 5 %. The sludges, primarily inorganic (e.g., hydroxide sludges), are mostly dewatered mechanically and disposed of in landfills.

Sludges with predominantly organic constituents may present serious problems in regulated disposal. Organic production sludges have properties comparable to those of municipal sewage sludges, but they are generally more voluminous and more difficult to dewater.

Sludges from metal-processing industries with high contents of heavy metals are regarded as hazardous waste. Much more problematical from the process engineering standpoint are many types of organic production sludges, because their dewatering or other required treatment is often difficult. In the case of fermentation residues, for example, conventional conditioning and dewatering processes rarely yield a sludge suitable for landfill disposal.

For a number of chiefly organic production sludges, the only available disposal method in the future will be thermal utilization after dewatering to a high solids content (and possibly drying). One example is sludges from papermaking. Sludges suitable for agricultural use, sometimes after intermediate composting, are chiefly those from food production. Dewatering is usually required here to lower transport costs.

7.6.6.2. Water Treatment Plant Sludges

Along with sludges from municipal sewage and industrial wastewater treatment plants, sludges generated in water treatment have begun to represent a disposal problem. In purely quantitative terms, these sludges have not yet become very significant.

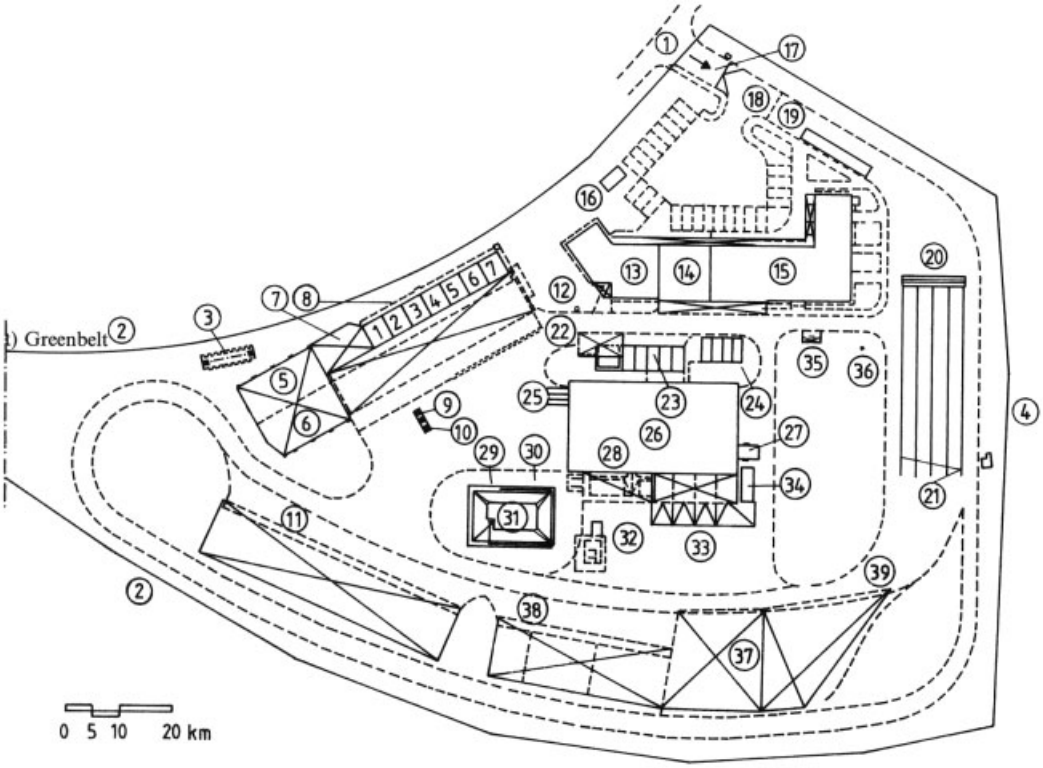

Figure 82. Intermediate storage site for hazardous wastes in Frankfurt/Main
1) Street; 2) Track system; 3) Buried steel tank for contaminated fire fighting water; 4) Sidewalk; 5) A1 storage; 6) Drum storage 3; 7) Equipment stand; 8) Sludge pits; 9) Silt trap; 10) Gasoline separator; 11) Stands for skips; 12) Boot washing station; 13) Social services building; 14) Spare part storage; 15) Operations building; 16) Substation; 17) Gate; 18) Barrier; 19) Scale; 20) Sampling pit; 21) Incoming truck parking; 22) Storage for toxic substances and samples awaiting analysis; 23) Storage; 24) Office containers; 25) Railway track; 26) Physicochemical treatment; 27) Ramp; 28) Chemical transfer; 29) Hydrant 1; 30) Lime bunker; 31) Storage basin; 32) Scale, 3 t; 33) Transfer station; 34) Sand separator; 35) Bicycle racks; 36) Hydrant 2; 37) Drum storage for solid wastes for incineration, recycling, and landfilling; 38) Drum storage for PCBs, acids, alkalies, and neutrals; 39) Equipment storage

More than 80 % of such sludges go to landfill disposal at present. Because they contain only small amounts of organic substances, thermal processes are not relevant. In the future, disposal of sludges generated in water treatment is likely to be monitored more closely, and criteria applicable to them will be similar to those for sewage sludges.

7.6.6.3. Dredging Sludges

Sludges produced in the dredging of waters represent a major potential problem.

In total, 100×10^6 m^3 of sludge is dredged annually from German waterways, seaports, inland harbors, lakes, and smaller bodies of water. Of this amount, some $(25-35) \times 10^6$ m^3 is piled on land; the rest is dumped elsewhere in inland waters or at sea.

Dredging sludges contain on average 60–65 % dry residue at the time of production, though much lower figures are also possible depending on the sludge composition.

For simplicity, the dredged solids can be regarded as half sand, in various grain-size distributions, half fines, whereby this distribution can vary to a wide extent locally. The organic content of the fine fraction varies widely and can be as high as 60 %.

Pollutants in the dredged solids accumulate chiefly in the fines. The concentration depends on the burden of pollutants entering the body

of water in question, but also increases with decreasing flow velocity. In rivers and streams particularly, the content of heavy metals often exceeds the limits set forth in the German Sewage Sludge Regulation.

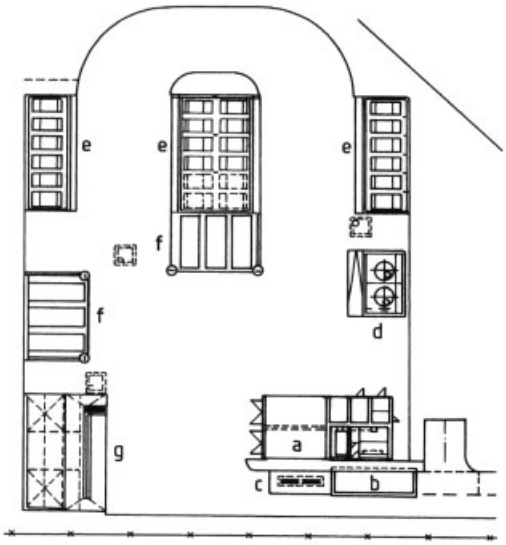

Figure 83. Intermediate storage site for hazardous wastes in Mitterteich
a) Operations building; b) Scale; c) Sampling pit; d) Storage tanks; e) Stands for skips; f) Semitrailer stands; g) Solids storage pile

Lakes have a significantly lower content of heavy metals, but the level of sapropel with its nitrogen compounds and phosphates as well as other oxygen-depleting substances is higher.

At inland sites in particular, the classical method of ponding is increasingly disallowed because it is no longer possible to find sufficiently large deposition areas near the shore, or because approval can no longer be obtained for a variety of reasons.

Even in coastal areas, drastic restrictions of this simple, cheap technology are becoming common.

Dredging sludges are commonly not considered for agricultural use because they often contain high levels of heavy metals and toxic organic compounds.

For the proper disposal and transportation of this material, it is necessary to dewater it to a high solids content and take account of the soil-mechanical aspects of landfill disposal.

The heavy-metal burden of dredged solids is largely associated with the fine fraction. If the coarse, sand fraction can be separated, the quantity of contaminated dredged material can be appreciably reduced.

The "traditional" method for separating sand from dredged solids is with the longitudinal-flow classifier, which exploits the difference in settling velocity of the sand and fines. The sand drops out in the fore part, near the inlet, while silt, clay, and organic substances pass on to the end, where the flow velocity declines so much that the fine particles can settle out. The main drawback to this method is the large area required.

Modern mechanical separation techniques are either still under testing or have been in full-scale use for only a few years. For example, after collection of coarse material in a hydrocyclone, the dredged solids can be divided into a sand and a fines fraction in a hydraulic classifier. The sand is washed and can be used as a building material. The mud goes to a centrifuge or other dewatering equipment, after which the solids can be disposed of.

The space requirement for such a processes is ca. 15 % of that for longitudinal-flow classification. Furthermore, the process can be run continuously, it is independent of weather, and there is virtually no seepage of contaminated water into the subsoil.

7.6.6.4. Sludges from Mining

Sludges originating from mining and quarrying generally exhibit good dewaterability and have a high solids content made up of granular particles. After dewatering, these sludges can be placed in landfills.

Restrictions may be applied in the future if the landfill criteria of the Technical Directive on Municipal Waste [266] cannot be formally complied with.

8. Intermediate Storage of Hazardous Wastes [272–275]

8.1. Introduction

Specially designed and equipped plants are required for the collection, intermediate storage,

processing, recycling, and disposal of hazardous wastes. These facilities must not only meet the requirements of relevant regulations and serve the ends of environmental protection but must also embody the state of the art.

For economic reasons, plants for the reutilization or disposal of wastes, such as incineration plants and landfills, are established on a supra-regional basis. It is therefore desirable, and sometimes also necessary, to erect intermediate storage sites, collecting points, and even small temporary storage facilities in order to ensure wide-area collection of hazardous wastes in accordance with industrial and commercial needs.

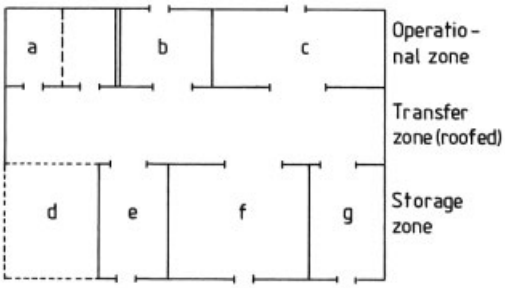

Figure 84. Intermediate storage facility for small quantities of hazardous wastes
a) Social rooms; b) Laboratory, office; c) Partly filled cases, working materials; d) Outdoor storage [intermediate bulk containers, containers (roofed)]; e) Firefighting zone III (wastes containing solvents); f) Firefighting zone I (chemicals, poisons, etc; batteries, mercury, PCBs); g) Firefighting zone II (compressed gas containers)

8.2. Intermediate Storage Sites

An intermediate storage site is a fixed plant in which wastes are received, held in pretreated form, and collected for forwarding to disposal sites.

Generally, intermediate storage sites consist of several zones, with a fairly consistent layout and usually the same equipment. The standard setup includes buildings with the receiving department and laboratory, a truck scale, and storage facilities for solid, liquid, and pasty wastes.

The main tasks of intermediate storage facilities is to reduce the volume of organically and inorganically polluted wastewaters that are subjected to physicochemical treatment. The latter generates detoxified, neutralized, and dewatered residues for recycling or landfill disposal; organic material for combustion; and cleaned wastewaters that can be released to a sewage works or receiving water.

Another function of intermediate hazardous-waste storage sites is the collection of small amounts of waste from medium-sized and small industrial operations, service areas, and households.

Figures 81 and 82 show two large intermediate storage sites in Munich and Frankfurt/ Main, both with physicochemical treatment systems. A smaller facility without physicochemical treatment is illustrated in Figure 83. Figure 84 shows how an intermediate storage site for small quantities of hazardous waste can be set up.

8.3. Zone Structure of an Intermediate Storage Site

An intermediate storage facility consists of several zones. For small units without processing systems, a site of ca. 5000 m^2 is sufficient. Large intermediate storage sites with treatment systems, such as physicochemical processing, may require 10 000 – 20 000 m^2.

The storage facility can generally be described in terms of the following zones: 1) Entry zone, 2) Traffic zone, 3) Storage zones, 4) Processing zone.

Entry Zone. The following are located in the entry zone:

1) Operations building with receiving, laboratory, and social rooms
2) Multipurpose building with storage rooms and workshop
3) Truck scale with holding area for arriving trucks waiting to be weighed
4) Sampling station with self-contained drainage and roof
5) Storage for samples awaiting analysis

The size of the operations building with its social rooms is governed by the number of employees. A small intermediate storage site can be run with three to five persons; a large one with a processing system such as a physicochemical treatment system would need a staff of 25 – 30. The laboratory, usually located in the

operations building, should be able to cover plant requirements and perform the monitoring and supervisory functions mandated by regulatory agencies. A high-performance, well-equipped laboratory can be a significant cost factor in an intermediate storage site.

The multipurpose building usually houses the machine shop, electrical shop, spare parts and equipment storage rooms, and space for on-site vehicles such as forklifts, suction trucks, sweepers, etc.

The truck scale is usually an 18 m scale; large facilities have two of these. The scale can weigh a complete tractor-trailer rig. Standard equipment includes a calibrated 3 t scale for weighing small quantities delivered in drums and casks.

Arriving trucks are briefly stopped at the sampling station (sampling pits) so that samples can be taken. The sampling pits have a self-contained drainage system and are equipped to protect the soil and groundwater. A roof allows falling rainwater to be drained off separately without becoming contaminated.

Traffic Zone. The traffic zone of an intermediate storage site is where hazardous wastes are transported to and from the storage and processing zones. The traffic zone must be impermeable to liquids and able to bear heavy loads. It must be large enough to allow vehicle operation without collisions and should offer free access to:

1) Rainwater and fire fighting water retention basins
2) Blind manholes and other inspection shafts
3) Onsite filling station for plant vehicles
4) Final pre-release inspection tank for wastewater
5) Car/truck wash
6) Substation and electrical distribution
7) Parking area

Storage Zones. Because hazardous wastes brought to the facility for intermediate storage come in a wide range of consistencies and are shipped in different kinds of containers, several storage zones have to be provided. Solids boxes or sludge pits are provided for solids, including pressed sludge filter cake, that arrive loose in containers or skips. The wastes are held and combined into larger units for shipping to the disposal site. As required by regulations, these storage areas are impermeable so as to protect soil and groundwater and are also roofed.

Another zone comprises roofed-over areas, isolated from the soil, where wastes delivered in skips and containers and destined to be held therein until further shipment can be set down and stored temporarily. Liquid and flammable hazardous wastes are commonly placed in separate tanks, located in a dike that retains leaks.

Hazardous wastes delivered in drums merit special attention. A drum storage area is provided where these drums and casks can be put in intermediate storage and combined into larger transport units. Generally, drum and cask storage zones are subdivided into sections for flammable and nonflammable substances as well as toxic and highly toxic substances. Casks and drums are held on pallets or in special trailers that can be used for forwarding.

In intermediate storage sites with treatment plants, a special zone must be kept free for wastes from transportation accidents, wastes generated by malfunctions, and wrongly declared wastes.

Treatment Zone. The treatment zone includes the organic and inorganic trains of the physicochemical treatment system with the appurtenant retaining basins and transfer equipment (see also Section 9.1). Separate systems are provided for organic wastes (wastes containing halogenated organic solvents, wastes containing nitrites, those containing cyanides, and miscellaneous organic wastes) and for inorganic wastes (alkaline wastes, wastes containing nitric acid and nitrates, those containing cyanides, nitrites, and chromates, and miscellaneous inorganic wastes).

The essential operations in the processing equipment are as follows:

1) Removal of coarse material
2) Separation of emulsions by chemical breaking, thermal treatment, and membrane techniques
3) Precipitation of metal salts
4) Sedimentation of sludges and slurries
5) Filtration of sludges in recessed-plate filter presses
6) Treatment of wastewaters (e.g., air stripping and activated- carbon filtration)

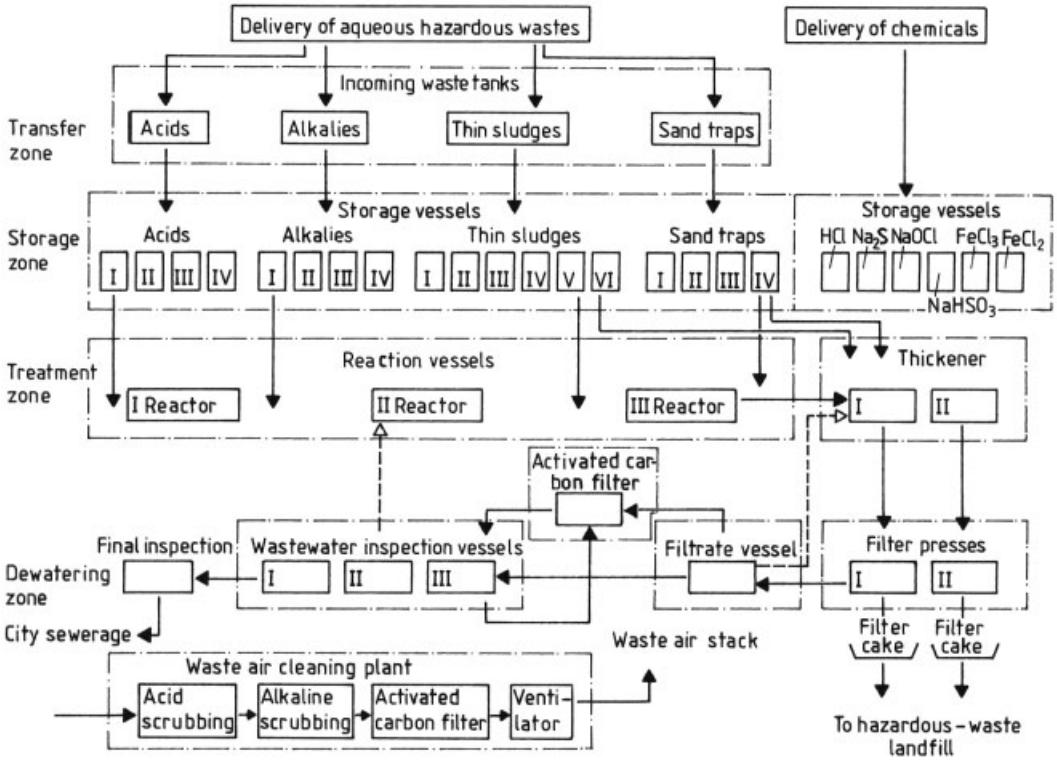

Figure 85. Schematic of the physicochemical treatment plant at Frankfurt/Main

7) Treatment of waste air, mainly in scrubbers and activated- carbon filters

Figure 85 shows of the physicochemical treatment system in the Frankfurt/Main intermediate storage site.

8.4. Procedures at an Intermediate Storage Site

An intermediate storage facility operates five days a week. Deliveries and transfers of wastes take place in the daytime; deliveries outside of normal receiving hours are possible only in emergencies. The physicochemical treatment plant runs on one or two shifts. Including plant management and laboratory personnel, around 25 employees perform all the work required in the intermediate storage site.

Wastes are delivered in tank trailers and trucks, skips, containers, and drums. Every delivery must be reported in advance by a waste disposal and recycling certificate. Before delivery is made, a schedule including quantity, type, and form of the hazardous waste is agreed on with the intermediate storage site. Manifests accompanying the delivery must identify producer of waste, transporter, and all substance-specific data for the wastes. Vehicles delivering hazardous waste are weighed on entry into the facility and their papers are examined to make sure they are properly completed. Next, a sample is taken from each load at the sampling station for incoming analysis in the laboratory. An internal load ticket is prepared with data from the first (declaration) analysis, which are compared with laboratory measurements and verified.

In the laboratory, the storage location, nature of processing and disposal, and disposal costs are set for all hazardous wastes received. Information about the unloading of wastes and the name of the responsible employee are entered on the internal ticket. This is the person who signs the receipt for the wastes and supervises

unloading. In this way, proper sorting and classification of wastes is ensured.

Drums and casks are subject to special handling, since improper handling of wastes, especially in drums, can have serious consequences. For this reason, every drum must have an adhesive label listing essential data such as the address of the waste producer, the type of waste, the container contents (quantity, composition, and consistency), and hazard warnings.

Drum labels are generally numbered in sequence, both with numerals and in barcode. All relevant data, such as drum number, internal load ticket, and so forth are collected with a data reader. These data are stored in the computer and the drum is placed in the drum storage area. In this way, a given container can be precisely located at any time. If the cask is later to be shipped away, for example in a trailer going to an incinerator, the data are forwarded electronically and the container is deleted at the intermediate site.

The recording, identification, and classification of all incoming hazardous wastes makes it possible to devise an overall logistical scheme. The dispatching office can then dispose the wastes according to how they can be processed or incinerated with allowance for the storage capacity. The chemical composition of the wastes is particularly important in this respect. The logistical system is important and helpful in, for example, incineration scheduling, where incineration and flue-gas cleaning must be adapted to the inlet pollutant levels. Furthermore, inclusion of all wastes in the logistical system makes it possible to report on the quantity, nature, composition, and current location of wastes at all times as required by the Industrial Emergencies Regulation (Störfallverordnung). Individual sites can thus be operated in accordance with waste composition and hazard potential, so that appropriate fire fighting measures can be employed in case of a fire.

8.5. Soil Protection

Stands and storage areas must be made suitably impermeable to prevent chlorinated and other halogenated hydrocarbons entering the soil and groundwater.

Chlorinated hydrocarbons have high vapor pressures and are relatively volatile, with relative vapor densities > 1, so that their vapors readily sink to the ground. The liquids are heavier than water, and are not readily ignitible. They are decomposed by the action of light and heat, yielding hydrochloric acid. The design and construction of soil protection measures must take all these properties of chlorinated hydrocarbons into account.

The type of soil protection depends on the degree and nature of use of the area in question. For example, storage zones with heavy traffic loads (e.g., skip and drum storage areas) must be made so that the traffic surface can withstand mechanical wear. As a rule, the protective arrangement here consists of two concrete courses with an intermediate layer of sheet steel or high-density polyethylene. The tightness of the surfaces must be verifiable, for example by vacuum testing, extraction and analysis of soil air, or leak monitoring.

Indoor areas meant to hold pollutant-laden liquids, such as drainage basins, tank complexes, self- contained drainage systems, and interceptor pipes, can be lined with poly(vinylidene fluoride) slabs, stoneware tiles impermeable to volatile halogenated hydrocarbons, or sheet steel. Again, the tightness of these surfaces must be verifiable.

Figures 86, 87, 88, 89, 90 show proven designs for soil protection.

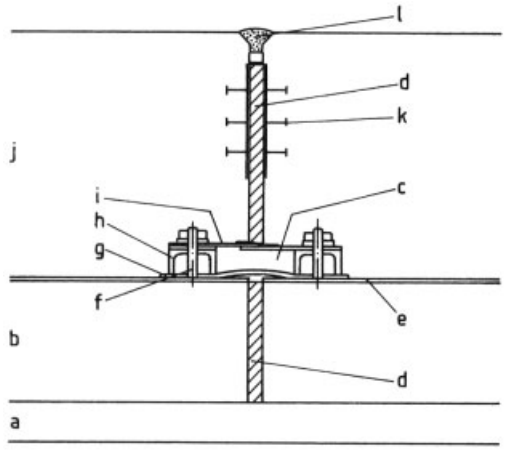

Figure 86. Steel liner and joint in soil protection system a) Improved subgrade; b) Reinforced concrete subbase; c) Expansion joint; d) Joint filler; e) Impervious layer (sheet steel); f) Bolt with tensioning nut; g) Fluorinated plastic seal tape; h) Channel section; i) Cover plate; j) Reinforced concrete protective layer; k) Joint sealing tape; l) Top seal

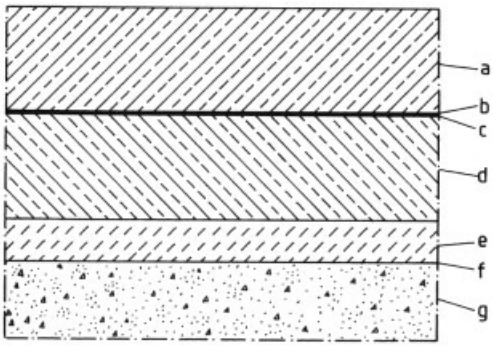

Figure 87. HDPE liner and joint in soil protection system
a) B35 reinforced concrete, $d = 25$ cm, crack-minimized, 0.2 mm; b) Surface pressure slab, $d = 0.5$ cm, made of HDPE; c) Delta barrier material (HDPE); d) B25 reinforced concrete, statically reinforced, $d = 25$ cm; e) Improved subgrade, B10 concrete, $d = 10$ cm; f) PE film; g) Subgrade

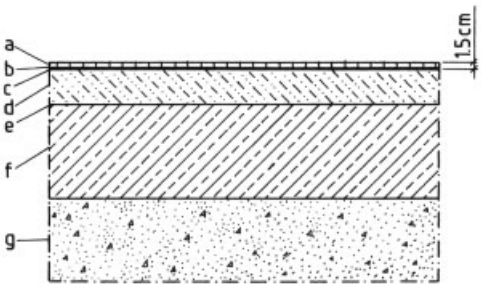

Figure 88. Crack-bridging coating in soil protection system
a) Keraplan EP-EL wearing course with priming; b) Keracid VEL-L coating; c) Kerapox leveling compound and Keranol EP1 priming; d) Cement floor with gradient, $d = 9.5$ cm; e) Plastic-modified adhesive bonding coat; f) B35 reinforced concrete, $d = 25$ cm; g) Subgrade

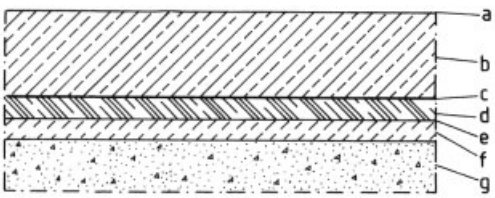

Figure 89. Steel lining in soil protection system
a) Keraplan-EG coating; b) Reinforced concrete slab, $d = 20$ cm; crack-minimized, max. crack width 0.2 mm; c) Sheet steel, welded, $d = 0.3$ cm; d) Retarded-set concrete, $d = 5$ cm; e) PE film; f) Lean concrete, $d = 5$ cm; g) Subgrade

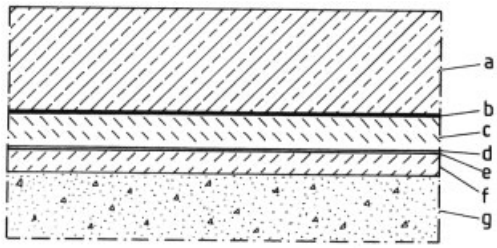

Figure 90. Film liner in soil protection system
a) B35 reinforced concrete, $d = 25$ cm; b) Sheet steel, $d = 0.6$ cm; c) B10 protective concrete, $d = 8$ cm; d) SLT slab (HDPE), $d = 0.25$ cm; e) Nonwoven mat; f) B10 protective concrete, $d = 6$ cm; g) Subgrade

8.6. Conclusions

Intermediate storage facilities have proved to be the best approach to wide-area collection of hazardous wastes. These sites, including those specialized for small quantities of hazardous wastes, are an important component of waste management. Proper sorting and pretreatment of hazardous wastes in intermediate sites can reduce transport quantities and distances and minimize danger to the public. Intermediate storage facilities must comply with all applicable laws, regulations, and codes and must keep up with the state of the art.

9. Treatment of Waste

9.1. Physicochemical Treatment

9.1.1. Introduction

The waste disposal chain ends at the landfill. Against the background of long-term landfill problems, it has become accepted that today's landfills must not become the contaminated sites of tomorrow. One consequence is that requirements on residues to be disposed of are becoming more stringent. Thus wastes intended for final disposal in landfills must not generate emissions that pollute groundwater, soil, or air.

When low-waste technologies [276] cannot be used to abate wastes and the wastes themselves cannot meet the criterion of "emission neutrality," thermal or physicochemical processes, or biological ones in some cases, must be employed for waste pretreatment.

This section deals with methods for the physicochemical treatment of wastes. The background for the discussion is the Second General Administrative Regulation pursuant to the German Federal Waste Act (Abfallgesetz), entitled " Technical Directive for Hazardous Waste" (TA Abfall) [277]. Among others, this regulation contains a complete list of requirements for stateof-the-art physicochemical processing plants. As the only statutory regulation embodying a detailed technical code, the TA Abfall has an influence reaching beyond Germany, serving as an acknowledged model within the European Union.

9.1.2. Types, Origin, and Properties of Wastes

To ensure consistent performance in waste disposal in the Federal Republic of Germany, the federal states in a cooperative body called Länderarbeitsgemeinschaft für Abfall (LAGA) [278] have prepared a "waste catalog" in which each waste is listed along with the treatment and disposal method that best meets environmental standards. The waste types of interest here are those for which the waste catalog prescribes physicochemical treatment. These fall into two groups:

1) Wastes with predominantly inorganic pollutants
2) Wastes with predominantly organic pollutants

Most such wastes have a liquid to pumpable consistency.

The first category includes alkalies, acids, electroplating sludges that have not been dewatered, concentrates, semiconcentrates, and rinsing waters from the chemical industry, from surface treatment and finishing (pickling and electroplating), and from the automotive and allied industries.

Wastes in which the pollutants are primarily organic include emulsions, wastewaters containing oil, synthetic coolants and lubricants, polymer emulsions, and organically loaded rinse and wash waters from the metalworking and automotive industries, from vessel and tank cleaning, from transportation facilities, and from military and comparable installations.

In both groups, the main component is water. Nonetheless, these wastes are not wastewater but hazardous wastes, because releasing them into sewerage without pretreatment is not permitted.

Inorganically contaminated wastes contain water-soluble compounds that are used in the above-mentioned industries. The cations are often nonferrous metals of groups 8 – 12 of the periodic table, such as copper, zinc, cadmium, and nickel. Toxic anions include chromate, dichromate, cyanide, and nitrite. Other typical wastes of this group are spent acids and alkalies. Along with chlorides and sulfates of the alkali and alkaline-earth elements, the burden of neutral salts often includes the corresponding ammonium compounds.

Complexing and chelating agents play a special role.

The main pollutants in the second group of wastes, with contaminants mainly organic in character, are aliphatic hydrocarbons such as those found in petroleum products. These can occur as "free" oil or dispersions and emulsions. In the latter case, the waste also contains emulsifiers such as anionic polyelectrolytes. Many other water-soluble organic products are also present, including surfactants, glycols, polar and nonpolar solvents (e.g., halogenated hydrocarbons), and aromatics with polar groups.

The total amount of water-soluble organic species is indicated by the chemical oxygen demand (COD), which ranges from a few thousand to a few hundred thousand milligrams of oxygen per liter.

Both groups of waste generally contain suspended solids of various particle sizes.

9.1.3. Treatment Processes

Physicochemical treatment processes have their origin in industrial wastewater treatment. Because of the complex composition of the wastes and the increasingly stringent requirements on the quality of liquid and solid residues generated in their treatment, this technology has continued to develop independently.

Many of the unit operations known from chemical technology are employed in physicochemical waste treatment in supra-regional hazardous waste disposal centers worldwide.

The techniques include mechanical, physical, chemical, and biological processes:

1) Mechanical processes: sedimentation, decantation, filtration (special cases: micro-, nano-, and ultrafiltration)
2) Physical processes: flotation, adsorption, ion exchange, distillation, stripping, reverse osmosis, evaporation, crystallization
3) Chemical processes: (leaching), neutralization, oxidation, reduction, precipitation, extraction
4) Biological processes: aerobic and anaerobic biodegradation, biological nitrate elimination

Not all these processes are equal in importance and scope of application. The physicochemical processes such as stripping/evaporation/crystallization [279] or reverse osmosis, and the biological processes are being used increasingly in waste treatment in response to rising standards.

Further impetus for physicochemical treatment has come from the development of methods for the purification of particular liquid wastes of complex composition. These wastes include leachates from municipal and hazardous waste landfills, which have high loads of both organic and inorganic species. In general, only a combination of methods – such as biological treatment, reverse osmosis, and evaporation/drying – can give a satisfactory result in such cases [280].

Progress in wet oxidation processes with H_2O_2 or ozone, utilizing catalysts or UV light [281], has been spurred by the same needs.

Electrochemical methods (cathodic deposition of nonferrous metals, anodic oxidation of cyanides and water-soluble organics) are of increasing interest. In future, because of the possible reclamation of substances from wastes, electrochemical processes combined with, e.g., solvent extraction [282] and electrodialysis will become more important.

Another process, developed in the United States and the United Kingdom, is the oxidation of dissolved and suspended, difficultly degradable organic compounds in supercritical water [283].

9.1.4. Requirements for Construction and Operation of Physicochemical Treatment Plants

TA Abfall includes provisions on process engineering, residues and wastewater, and instrumentation and control in physicochemical treatment plants. The goal of physicochemical treatment, as defined in this regulation, is that the treated waste be rendered recyclable or disposable by a downstream operation such as incineration or landfilling.

Treatment Plan. Prior to industrial-scale treatment of a waste, a treatment plan must be prepared on the basis of laboratory tests. Wastes delivered to public physicochemical treatment facilities are commonly single batches of multiply and severely contaminated material that the generating plant cannot handle in its own wastewater treatment system because of expected complications (e.g., the presence of complexing agents).

The waste disposal operator must therefore conduct lab trials to identify the treatment steps for each batch. For example, the treatment chemicals and their quantities must be determined to sufficient accuracy and passed on as instructions to the operating personnel.

Organic and Inorganic Treatment Trains. The "treatment trains" for organically and inorganically contaminated wastes should be separate because they differ in their chemical and physical properties. This also avoids mixing of wastes and makes it possible to meet the differing requirements for the disposal of the resulting solid and liquid treatment residues. The following processes are considered state-of-the-art for the organic treatment train: distillation, chemical emulsion breaking, and membrane processes.

In addition to these basic processes, the overall process must also include other unit operations. For example, the proper treatment is often preceded by the separation of solid and liquid contaminants. This holds equally for the organic and inorganic trains.

A variety of total concepts are under discussion for the treatment of inorganic wastes, and a distinction is made between monofunctional and multifunctional utilization of plant sections.

In a monofunctional plant, only one quite definite treatment, such as cyanide oxidation, can be performed. A multifunctional plant design, in contrast, makes it possible to interconnect the plant components flexibly, so that a variety of reactions can be carried in the same portion of the plant.

Precipitation of heavy metals is a point of central importance in inorganic treatment, because of their high water-pollution potential.

Dissolved heavy metals are first precipitated as hydroxides or hydrated oxides. If hydroxide precipitation does not meet the required limit, a second stage, such as sulfide precipitation, is added. Sulfide precipitation alone is not permitted due to the toxicity of hydrogen sulfide and the possibility of remobilizing processes acting on heavy metal sulfides in the landfill.

Posttreatment. Classical treatment processes are often insufficient for ensuring compliance with present emission limits. For such cases TA Abfall specifies possible posttreatment steps. Supplementary treatment processes downstream of the organic train include stripping, activated- carbon adsorption, and biological treatment; downstream of the inorganic train, reverse osmosis and evaporation.

Instrumentation and Control. All processes involved in physicochemical treatment require monitoring. This is especially true of processes generating potentially toxic substances as intermediates or byproducts. The degree of automation of these operations has to be determined case by case; it depends on the composition of the inlet wastes, among other factors. As a rule, physicochemical treatment plants run by hazardous waste disposal companies are operated in batch mode. This is how they differ from industrial wastewater treatment systems, which normally are operated in continuous (flow-through) mode.

9.1.5. Technological Concept of Physicochemical Treatment Plants

A comparison of existing physicochemical treatment systems operated by the major hazardous-waste disposal companies shows that the process engineering is quite similar from one to another. For example, sludges from neutralization and hydroxide precipitation are always dewatered in recessed-plate filter presses.

9.1.5.1. Treatment Plant for Inorganically Contaminated Wastes

Figure 91 presents the flowsheet of a conventional physicochemical treatment plant for waste with predominantly inorganic pollutants.

The plant consists of the following units:

1) Separate receiving and holding basins for acid wastes, alkaline wastes containing cyanides and nitrites, wastes containing chromates, and neutral sludges that have not been dewatered
2) Chemical tanks for acids, alkalies, and other treatment chemicals such as sodium hypochlorite, sodium hydrogensulfite, hydrogen peroxide, and flocculation aids
3) Reaction vessels of multifunctional design in which any treatment operation can be performed in any vessel
4) Dewatering (thickener, sludge pump, and recessed-plate filter press)
5) Final wastewater tanks, which are filled, then inspected by the laboratory, and finally discharged

This physicochemical treatment system is a widely used classical solution. The technical provisions of the Technical Directive on Hazardous Waste (TA Sonderabfall), as well as wastewater standards now in force or being prepared, require a more complex plant configuration.

An advanced concept, essentially oriented to the TA Sonderabfall provisions, includes the following:

1) Twelve receiving stations, for example, stations for neutral sludges; acids; alkalies; sludges containing cyanides, chromates, nitrites, and complexing agents; acid wastes containing HNO_3; etc.
2) An appropriate number of vessels for intermediate waste storage
3) A chemical supply system

An advanced treatment section includes four independent pretreatment lines ("main detoxification"):

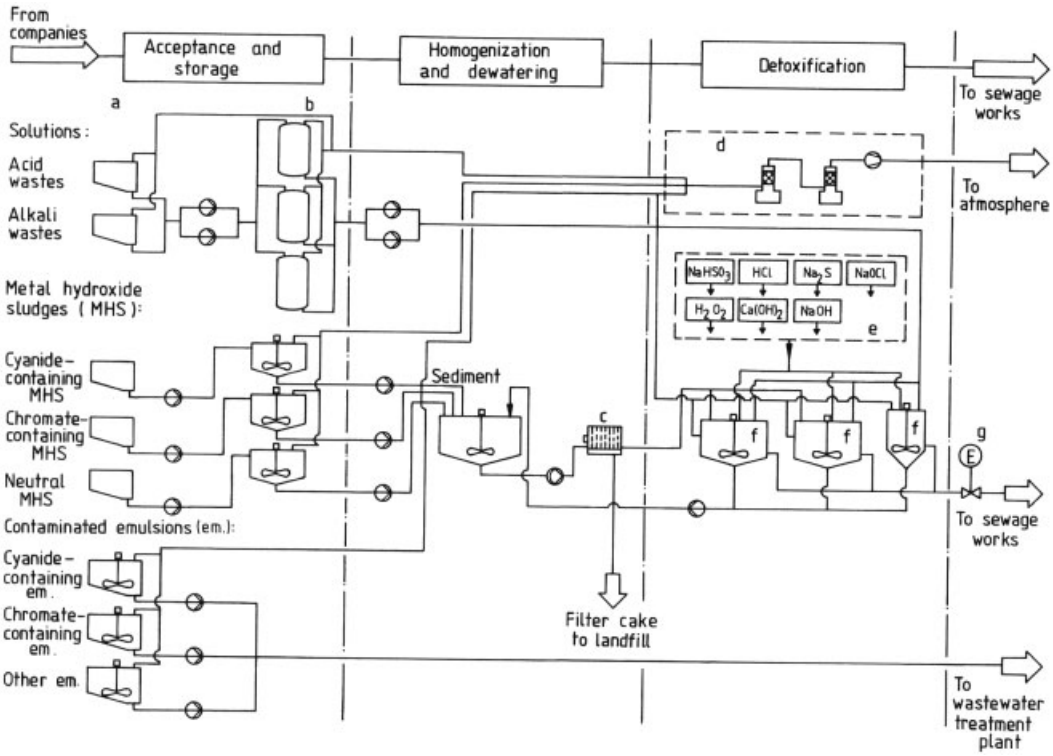

Figure 91. Conventional physicochemical treatment system for wastes containing inorganic contaminants
a) Incoming waste basins; b) Holding basins; c) Recessed-plate filter press; d) Off-gas cleaning unit; e) Chemicals station; f) Treatment vessels; g) Emergency valve

1) Metal hydroxide and neutral sludges as well as acid wastes
2) Sludges containing cyanides, chromates, and nitrites; alkaline sludges
3) Complexing agents
4) Emulsions containing cyanides, nitrites, and other inorganic contaminants

Discharges from the first two treatment lines pass via buffer storage to a common final treatment step, operating in continuous mode. After sulfate precipitation and alkalization, the operations performed here include AOX/NH_4 stripping, multiple-bed filtration (to remove fine solids), selective cation-exchange treatment (to remove heavy metals), UV/H_2O_2 treatment (to remove organic species), and filtration with activated carbon.

Wastes containing complexing agents have their own separate treatment system. Contaminated emulsions are subjected to preliminary detoxification and then fed to an independent physicochemical treatment unit.

9.1.5.2. Treatment Plant for Organically Contaminated Wastes

Figure 92 shows the flowsheet of a conventional treatment system for organically contaminated waste of liquid to pumpable consistency. The major difference between processes in existing plants is the method used for emulsion breaking. Techniques employed are chemical salt separation (precipitation/flocculation), thermal breaking (evaporation), and membrane processes (ultrafiltration).

The process generally includes the following steps:

1) Acceptance

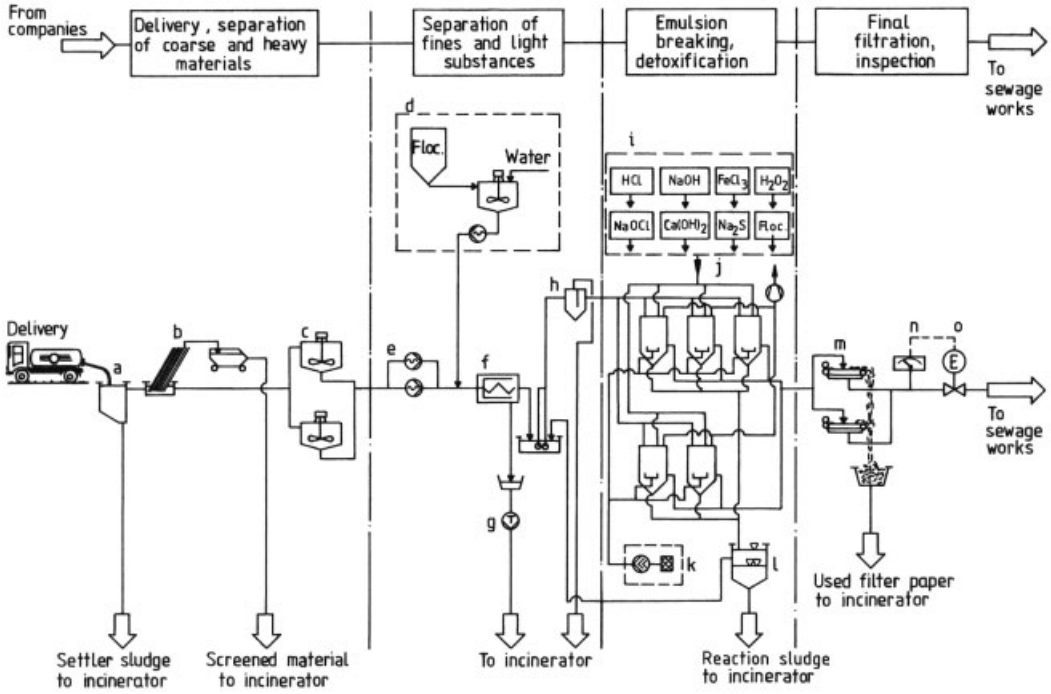

Figure 92. Conventional physicochemical treatment plant for wastes containing organic contaminants
a) Settling basin (four chambers); b) Screens; c) Holding basins; d) Flocculant metering station; e) Pumping station; f) Decanter; g) Sludge pump; h) Oil separator; i) Chemicals station; j) Reaction vessels; k) Compressed air station; l) Sludge thickener; m) Paper belt filters; n) Flow measurement; o) Emergency valve

2) Separation of solids by sedimentation, supplemented by mechanical separation (e.g., decanter or centrifuge) if required
3) Separation of light oversize material and free oil
4) Holding for homogenization
5) Emulsion breaking
6) Reaction sludge separation (only when salt separation is used in conjunction with precipitation/flocculation) by flotation, filtration, or centrifugation
7) Wastewater holding for final inspection before discharge

The flowsheet of a modern plant, complying with TA Abfall, for treating organically contaminated hazardous wastes is shown in Figure 93. This system features improvements in acceptance and pretreatment of wastes, with modified emulsion breaking, efficient waste air treatment, and equipment for the removal of volatile and nonvolatile halogenated hydrocarbons.

The principal advance over conventional systems, however, is in the posttreatment of the wastewater. In the example shown, denitrification is followed by two-stage aerobic biological treatment. Each of the two activated sludge stages includes an activated sludge tank and a settling tank.

Results from the operation of one such plant [284] show a 60 – 90 % reduction in COD and > 90 % reduction in BOD_5.

9.1.6. Environmental Requirements

Residues. Processing of inorganically contaminated wastes generates metal-containing hydroxide sludges in the form of filter cakes, dewatered in filter presses. These filter cakes must be deposited in hazardous waste landfills unless facilities exist for processing them to reclaim their nonferrous metal content.

Physicochemical treatment plants for organically contaminated wastes must be integrated with hazardous-waste incineration plants for

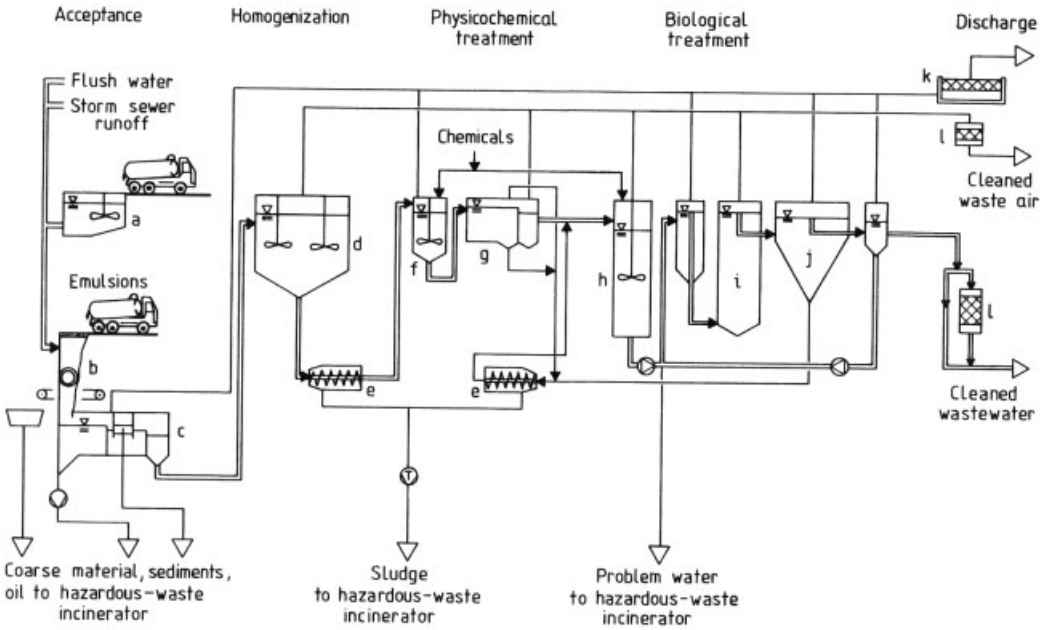

Figure 93. Advanced physicochemical treatment system with final biological treatment for wastes with organic contaminants a) Preliminary detoxification basin; b) Rotary screen; c) Incoming waste basin; d) Holding basin; e) Sludge separation; f) Reaction vessel; g) Flotation; h) Supply vessel; i) Activated sludge tank; j) Secondary settling tank; k) Biological waste-air filter; l) Activated carbon filter

disposing of the oil-containing solid and liquid residues produced.

Wastewater. The treated wastewater discharged as the end product of physicochemical treatment is stored in final inspection tanks before being released. Complete laboratory tests must be performed to ensure that regulatory standards are met.

Physicochemical treatment plants of conventional design (Figs. 91 and 92) were generally able to satisfy legislative requirements on wastewater quality up to the mid-1980s. Standards have since become much more stringent as a result of new legislative and regulatory provisions, and this development is expected to continue in the short term. Examples of new statutory regulations can be found in the amended Water Resources Management Act (Wasserhaushaltsgesetz, WHG) of the Federal Republic of Germany adopted on September 23, 1986; to date, more than 50 general wastewater regulations have been issued pursuant to § 7a of WHG, establishing minimum standards for wastewater quality in various industries. A draft regulation entitled "Treatment of Production-Specific Residues and Product Residues" will set forth higher standards for the quality of waste-water from physicochemical treatment plants. For selected parameters, Table 18 gives limits dating from the 1970s, corresponding values from the late 1980s, and anticipated values taken from the draft regulation.

Waste Air. When organically and inorganically contaminated wastes undergo physicochemical treatment, waste air contaminated with pollutants is produced. In the first case, the off-gas contains volatile organic substances such as short- chain aliphatic and aromatic hydrocarbons, volatile halogenated hydrocarbons, and odorous nitrogen and sulfur compounds.

Waste air is generally treated with activated carbon, which in turn has to be disposed of as waste after use. When activated carbon is used, displacement reactions between species that compete for adsorption sites must be considered. To prevent breakthrough of pollutants, the activated carbon must be replaced before it is exhausted.

Table 18. Development of emission limits for selected substances from physicochemical treatment plants

Parameter	Unit	1979 limits	1989 limits	Draft limits 1992
Cadmium	mg/L	3.0	0.1	0.2
Chromium (VI)	mg/L	0.5	0.1	0.1
Copper	mg/L	1.0	0.5	0.5
Nickel	mg/L	2.0	0.5	1.0
Ammonium	mg/L		200	10.0 (NH_4-N)
Nitrite	mg/L	30	5.0	2 (NO_2-N)
Nitrate	mg/L			20 (NO_3-N)
COD	mg O_2/L		900	200
Hydrocarbons	mg/L	5.0	10.0	2.0
AOX	mg/L		1.0	1.0

Odorous substances present in low concentrations, if they are biodegradable, can also be removed on biofilters. An air treatment system for an organic physicochemical treatment plant is reported in [285]. It splits the air stream into a smaller stream with a high burden of hydrocarbons and volatile halogenated hydrocarbons, which is led through an activated carbon filter, and a stream about ten times larger in volume, which mainly contains odorous substances. This larger stream can be effectively cleaned by passing it through a biofilter (area 200 m^2, bed depth 1 m).

Exhaust air from inorganic physicochemical treatment systems contains volatile reaction products and intermediates, such as nitrogen oxides, acid-containing aerosols, and ammonia, but also volatile organic compounds. These substances are removed in a packed-tower wet scrubber, sometimes combined with activated carbon filters.

9.1.7. Storage and Transportation

Storage. Among the provisions dealing with the storage of liquid wastes is the "Regulation on Facilities for the Storage, Filling, and Transfer of Water-Polluting Substances" (Verordnung über Anlagen zum Lagern, Abfüllen und Umschlagen wassergefährdender Stoffe, VAwS) [286].

A fundamental requirement is chemically appropriate layout of tanks. For example, the positioning of a storage tank for cyanide-containing wastes next to a vessel for acid wastes is not acceptable.

Tanks must be placed in catch trays, which must be made of materials resistant to the tank contents. When the presence of halogenated hydrocarbons cannot be ruled out, the authorities require monitorable floor protection resistant to volatile halogenated hydrocarbons (e.g., a steel liner).

The tanks must be made of resistant materials. Steel vessels with corrosion-resistant coatings and tanks made of polyethylene or polypropylene have proved suitable.

Water quality law in the Federal Republic of Germany generally requires design approval for storage tanks; for special designs, approval by an authorized agency such as the Technischer Überwachungsverein (TÜV) is necessary.

Because liquid wastes generally contain suspended solids, tanks must have conical bottoms and stirrers to ensure homogenization.

Because of the hazard potential of wastes held in temporary storage, regulations such as TA Abfall set strict requirements on tank instrumentation and monitoring. Level measurement is required in every case, together with safeguards against overfilling. Temperature and pressure monitors may be required on a case-by-case basis.

Because the tanks must be cleaned regularly to remove encrustations and deposits, they must be provided with access and inspection holes.

Transportation of liquid wastes from the waste-producing site to the disposal site is subject to strict regulations.

Suitable vehicles, such as chemical-resistant suction vehicles, are required for shipping large quantities (5 – 12 m^3). Small quantities may be transported only in safe containers, such as approved-design cubic tank containers or drums.

As a rule, only duly licensed trucking firms may transport such wastes. To obtain regulatory approval as a carrier, a company must demonstrate the technical suitability of its fleet and the technical knowledge and skill of its personnel.

The hazardous properties of wastes require that proper occupational health and safety practices must be in effect wherever employees may be exposed to wastes while handling them (draining, filling, sampling). This ranges from protective clothing, gloves, spectacles or goggles, and protective masks to coveralls and full respiratory protection. Personnel working in treatment plants are subject to continuous medical surveillance.

9.1.8. Legal and Economic Aspects

The legal framework is that of environmental law. Along with waste management law and the related regulations (e.g., TA Abfall and the Waste Transportation Regulation (Abfallbeförderungsverordnung) mentioned above), applicable fields of law include pollution control (environmental compartment: air), water quality (environmental compartment : water), and numerous administrative and other regulations and technical guidelines.

In addition to national law and, in part, superior to it, there are also a number of EC directives either in force or being prepared to create unified standards for the treatment of hazardous wastes in the European Union.

The large number of legislative provisions and stringent technical requirements for environmental protection have their price. Physicochemical treatment of hazardous wastes is a cost-intensive operation. Disposal costs include the high capital costs of the plant and depend on the type and concentration of the pollutants being treated.

Another factor contributing to the high costs is the disposal of the solid residues generated in physicochemical treatment. These residues require hazardous waste landfills or hazardous waste incineration plants. In 1993, the cost of physicochemical treatment was 150 – 1500 €/t, with an average of ca. 260 €/t.

9.2. Hydrolysis and Alcoholysis of Plastics Wastes

9.2.1. Introduction

Because of the growing amount of waste being produced and increasing environmental consciousness, recovery of raw materials from wastes is becoming of increasing importance. In 1989, with an output of ca. 9×10^6 t of plastics, the Federal Republic of Germany generated ca. 2.2×10^6 t of plastics wastes [287]. Of this, ca. 500 000 t of sorted and 15 000 t of mixed residential waste was recycled. In 1992 ca. 45×10^6 t of plastics were used in Germany. These quantities show how important recycling of plastics wastes is going to be in the future. The thermal and chemical processes now used for plastics recycling can be classified in two categories [288]: 1) energetic utilization (combustion) and 2) material reclamation by chemical reaction (pyrolysis, hydrolysis, hydrogenation). The processes that have so far found large-scale application have been those involving energy utilization. A pilot plant operated at Bottrop by VEBA Öl is hydrogenating plastics wastes to obtain a naphtha-like mixture of hydrocarbons. Other pilot and demonstration facilities are in the planning stage.

Hydrolysis and alcoholysis processes for recovery of feedstocks from plastics wastes have not yet been studied in depth. This section therefore gives a summary of scientific and application-oriented research reported in the specialist and patent literature, as of well as planned processes.

9.2.2. State of the Art

For a plastic to be broken down into its monomeric constituents by hydrolysis or alcoholysis, its polymer structure must contain polar groups (ester, amide, urethane, carbonate, or acetal) along with the hydrocarbon groups. Such a polymer can be decomposed under relatively mild conditions with the use of various additives. Aminolysis is of secondary importance and is not discussed here.

9.2.2.1. Hydrolysis of Polyurethanes

Polyurethanes are bulky plastics (density ca. 30 kg/m^3, storage volume ca. 35 m^3/t). The environmental and economic problems linked with their high bulk volume, plus the relatively high manufacturing cost (13 – 15 €/kg), make an economical reprocessing method necessary. Polyurethanes are unstable when heated, but thermal processing yields numerous degradation products, which are so diverse that they cannot be reclaimed. However, the use of selected chemical recycling methods offers some advantages. The hydrolysis of polyurethanes is carried out at high temperatures and pressures to increase the reaction rate [288–291]. Reaction products include CO_2 and the polyether alcohols, polyester alcohols, and amines used as feedstocks in the original plastics production process (Scheme 1).

Hydrolysis Plant. Processes reported in the literature are of batch [289], [290] and continuous extruder types [291]. Figure 94 shows a flowsheet for a continuous pilot plant hydrolyzing flexible polyurethane foams [287], [291]. The foam, produced from toluene diisocyanate and polyols, is decomposed, whereby the isocyanate portion is broken down into the original toluenediamine (TDA), of which 90 % is recovered.

The decomposition of polyurethane in the presence of water is completed within ca. 2 h at 473 K; if the reaction temperature is increased by 10 – 15 K, the residence time required for complete reaction is halved. However, if the temperature is too high, the monomers obtained are not well-defined and, hence, not reusable.

The continuous process of Figure 94 is made possible by a twin-screw extruder that generates and maintains a constant steam pressure of 15 – 30 bar at reaction temperatures of 473 – 503 K. The PUR foam wastes, reduced to flakes or powder, enter a hydrolysis screw continuously via a metering screw. The hydrolysis screw operates at up to 573 K and 100 bar. Physical degassing of the plastic takes place in the compression zone. The pitch of the screw flight in the barrel decreases stepwise, giving a high degree of compression. In this way, the barrel is divided into three compression sections plus the reaction zone, which is separated from the upstream portion by a left-hand screw. Water required for polyurethane hydrolysis is injected by metering pumps and is vaporized in the extruder. Intensive mixing of water and plastic is promoted by kneader elements. The mean residence time of the mixture in the reaction space is ca. 20 min. The resulting mixture of gas (vapor) and liquid is continuously delivered to the reprocessing apparatus via a pressure release device. Volatiles entrained by product CO_2 and other gases can be recovered by condensation in a downstream cooler. Up to 90 % of toluenediamine is isolated from the product mixture in a complicated process involving aqueous acids. Integrated distillation and extraction stages are then used to separate all components. This is complicated by the emulsifiers from the polyurethane foam.

Separation of Hydrolysis Products – Physical Properties of Recycle Stream. The polyols can be recovered in pure form by reacting the crude hydrolysis product with oxalic acid. Oxalic acid and the resulting oxalate salt of TDA are only sparingly soluble in the pure polyol. The salt is removed by filtration, and adhering polyether is washed off with toluene. The toluene is distilled off to give an amine-free polyol. Alternatively, HCl gas can be admitted directly into the hydrolysis product. Under suitable reaction conditions (343 – 364 K, 50 % dilution of hydrolysis product with toluene), formation of TDA monohydrochloride is favored; this product can be recovered by filtration [291], [292].

The choice of separation process is dictated by the plastics waste being processed [291], [292]. The key parameters are chain length, degree of branching of the polyol, and hydrophilic character, which depends on the number of primary OH groups in the polyol. The process becomes more difficult if the PUR has been modified. For example, if it contains amine, the hydrolysis product derived from the standard flexible slab stock recipe is strongly alkaline. Hydrolysis of flame retardants containing phosphorus and chlorine yields phosphoric and hydrochloric acids, which form salts with the amines. If the mole fraction of acid-forming substances is greater than the amine content, the hydrolysis solution becomes acidic, and all the amines are present as salts; this complicates amine recovery. Amine yields are 85 – 92 %, depending on

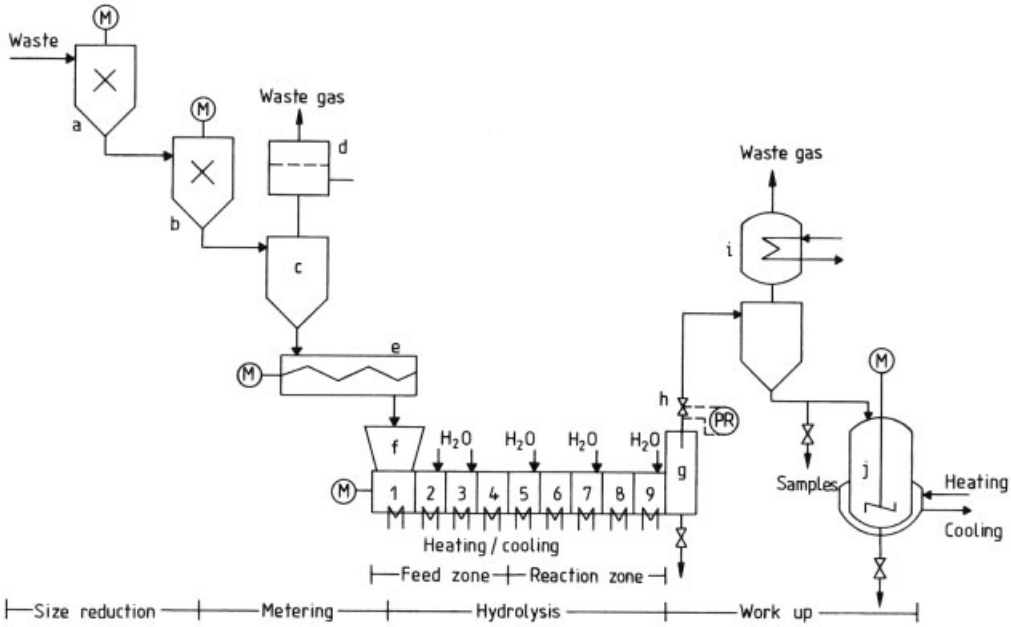

Figure 94. Flowsheet of a continuous plant for hydrolysis of polyurethanes
a) Alpine mill; b) Impeller mill; c) Hopper; d) Filter; e) Metering device; f) Feed hopper; g) Discharge section; h) Control valve; i) Cooler; j) Holding tank

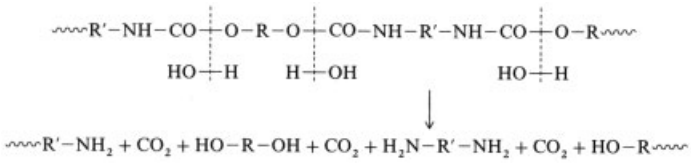

Scheme 1. Hydrolysis of polyurethanes

the waste inlet to the process; polyol yields, in contrast, are largely independent of recipe and range between 95 and 98 %.

The quality of the recovered polyol is summarized in Tables 19, 20, 21 [291], [292]; oxalic acid was used to separate TDA. Table 19 compares the properties of a Bayer grade of pure polyether polyol with those of recycled polyol. The various compositions of standard flexible foams, obtained by manual foaming of mixtures of standard and recycled polyether polyols with toluene diisocyanate, can be found in Table 20. Table 21 describes the properties of the foamed products as well as their foaming behavior.

Table 19. Properties of polyether polyols

	Specification Desmophen 7100	Regenerated polyether (average values)
Hydroxyl number	49 ± 3	50.9
Acid number	< 0.1	0.63
Water content, %	< 0.1	0.065
Viscosity at 25 °C, mPa · s	520 ± 30	575
Refractive index	1.4542	1.4551
Ash, %	< 0.05	0.05
Unsaturated compounds, meq./g	< 0.05	0.095
Peroxide content, ppm	< 30	~ 5
pH	6.6 to 7.2	6.0
Solidification point, °C	< -40	< -50

Table 20. Foam formulations for standard and regenerated polyether polyols

Trial	Standard		H_2O	Catalyst			TDI ∗
	Polyether polyol	Regenerated polyether polyol		A	B	C	
1	100		4.0	0.5	1.0	0.25	50
2	90	10	4.0	0.5	1.0	0.2	50
3	80	20	4.0	0.5	1.0	0.1	50
4	50	50	4.0	1.0	0.8	0.05	50
5		100	4.0	1.2	0.6	0.2	50
6		100	4.0	1.2	0.6	0.25	50

∗ TDI = toluene diisocyanate.

Table 21. Foaming behavior and foamed plastics properties

Trial	Initiation time, s	Rise time, s	Setting time, s	Air permeability, mm H_2O	Bulk density [a] kg/m^3	Tensile strength, kN/m^2	Fracture strain [b] %	Compression hardness at 40 % compression [c] kN/m^2
1	9	86	28	150	27	110	190	4.1
2	6	82	41	160	25/26	105	170	4.0/3.9
3	7	113	84	210	26/26	95	155	3.9/3.9
4	7	83	91	220	27/27	85	135	3.8/3.9
5	6	54	61	250	23/23	95	180	3.7/3.7
6	5	57	54	230	25/25	105	190	4.0/4.1

[a] DIN 53420;
[b] DIN 53571;
[c] DIN 53577.

Table 22. Properties of regenerated and standard polyether polyols as a function of the separation process [292]

	Regenerated polyether polyol				Standard
	HCl Separation ∗			Oxalic acid separation	Desmophen 7100
	A	B	C		
Hydroxyl number	48.5	45.9	47.2	50.9	49 ± 3
Acid number	0.77	0.02	0.16	0.63	< 0.1
Water content, %	0.1	0.06	0.075	0.065	< 0.1
Viscosity at 25 °C, mPa · s		570		575	520 ± 30
Refractive index		1.4539		1.4551	1.4542
Ash, %		< 0.05		0.05	< 0.05
Unsaturated compounds, meq./g		0.117		0.095	< 0.05
Color (Gardner)	ca. 5	3 – 6	4 – 5	7 – 15	< 1
pH	5.0	8.2	6.7	6.0	6.6 – 7.2

∗ A: two-stage HCl precipitation without post-treatment; B: post-treatment with solid, anhydrous soda; C: neutralization with 10 % KOH.

The data of Tables 19, 20, 21 show that technically acceptable products are obtained when recycled polyol is used. However, these results were obtained in manual foaming tests, and semi-industrial scale production processes must be used for a final assessment. Higher-quality recycled polyols can be obtained by HCl separation of TDA. Table 22 lists the properties of polyols as a function of the separation process and compares them with the standard product [292]. The low tinting values and improved color stability of polyols after HCl separation are especially promising for industrial reuse.

In addition to the tests on flexible slab stocks, flexible molded foams were broken down into

their chemical constituents so that the effect of formulation on both hydrolysis conditions and regenerate quality could be studied. The reaction conditions found for hydrolysis and TDA separation (HCl method) on flexible slab stocks were also applicable to this class of foamed plastics. Table 23 compares properties of the regenerated polyether polyols with those of a standard product (Desmophen 3900).

Like the commercial polyol Desmophen 3900, the regenerated polyether polyol exhibits good foaming properties. The molded foam showed tolerable deviations from commercial foams.

In conclusion, it must be stated that the hydrolytic reprocessing of plastics wastes has at present no prospects for development into a marketable process for economic reasons. Much more interesting are the alcoholysis processes, in which a variety of alcohols are used to decompose the plastics wastes into their basic components.

9.2.2.2. Alcoholysis of Polyurethanes

Unlike hydrolysis, the alcoholysis of polyurethanes does not generate CO_2 because the reaction products are more stable (Scheme 2) [287].

This reaction yields a mixture of the starting polyols and low molecular mass urethanes with ether terminal groups. If a diol is used for polyurethane alcoholysis, the reaction products have terminal OH groups and can thus be directly reacted with isocyanates. The use of special alcohols or diols in this degradation reaction makes it possible, during the actual recycling process, to match the properties of the product mixture to the intended application.

Extensive work on this recycling process has been reported in the literature [287]. Topics of study have included, in particular, the effect of different alcohols and catalysts on the product distribution. Higher-boiling glycols based on ethylene oxide and propylene oxide have proved suitable. With these cheap substances, used alone or mixed, the reactivity and the properties of the regenerated polyols can be controlled within certain ranges. Basic substances (amines, hydroxides, alkoxides) and Lewis acids catalyze this reaction.

Ford Motor Company has reported a combination of glycolysis and hydrolysis, called hydroglycolysis [293], [294]. If pure diethylene glycol is used to treat PUR wastes, a mixture of widely varying reaction products is obtained. The complex product distribution makes reprocessing difficult. The product spectrum is greatly simplified if small amounts of water (ca. 0.1 – 2.4 wt %) and basic catalysts are added to the reaction mixture. The only primary products in this case are toluenediamine and polyether triol. LiOH has been found to be a good catalyst for the reaction; NaOH, KOH, $Ca(OH)_2$, $Ba(OH)_2$, $Mg(OH)_2$, and $Al(OH)_3$ proved, however, unsuitable.

The polyol obtained by alcoholysis can be recovered in high purity by liquid – liquid extraction with high-boiling hydrocarbons such as hexadecane. The regenerated polyol can be reused in polyurethane synthesis after mixing with new polyol (1 : 1 ratio). A plant designed by Ford would be profitable according to cost estimates based on 1982 U.S. prices, if there is no opportunity to sell the PUR waste and the regenerated polyether, polyol can be returned directly to the production process.

In 1993, REGRA Recycling used alcoholysis of polyurethanes as the basis for developing a process suitable for total recycling of all PUR wastes generated in the production and processing of these plastics [295]. The regenerated polyol can be returned directly to the production process. The PUR wastes are ground and fed to an unpressurized reactor where the temperature is raised and alcohols and other unspecified additives are added. The polyurethane wastes are broken down into fragments of various size by chemical and physical processes. The product distribution (type of polyol) can be controlled by means of the reaction conditions (temperature, alcohol, catalyst). The filtered regenerated polyol can be held in supply tanks until it is reused.

Table 24 compares the physical properties of integral polyurethane foams, one made from pure polyol and the other with the addition of regenerated polyol. As in the hydrolysis case, there is no significant difference between the products.

Table 23. Properties of recycled polyether polyols obtained by HCl separation compared with a standard product [292]

	Recycled polyether polyol ∗		Desmophen
	A	B	3900
Hydroxyl number	37.6	36.8	35 ± 3
Acid number	0.05	0.4	< 0.1
Water content, %	0.05	0.08	< 0.1
Viscosity at 25 °C, mPa · s	830	840	825 ± 50
Refractive index	1.4541		1.4450 ± 0.002
Ash, %	< 0.05		< 0.2
Unsaturated compounds, meq./g	0.064		0.05 – 0.06
pH	9.3	5.8	6.5 – 7.5
Color (Gardner)	14 – 15	15 – 20	< 1
Saponification number	0.89		< 0.1
Peroxide, ppm	7		< 10
Aldehyde, %	0.003		< 0.0005
Ionol ∗∗, %	0.09		0.23 – 0.25
Toluene, %	0.16		< 0.15
Clear point, °C	48		49 ± 1
NH_2 content, %	0.03	0.03	
Solidification point, °C	− 19.3		−38

∗ A: two-stage HCl precipitation with soda treatment; B: two-stage HCl precipitation without soda treatment.
∗∗ Ionol = 2,6-di-*tert*-butyl-4-methylphenol

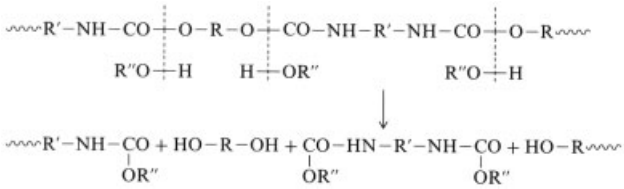

Scheme 2. Alcoholysis of polyurethanes

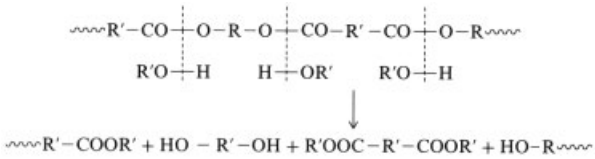

Scheme 3. Alcoholysis of polyesters

Table 24. Physical properties of semirigid integral polyurethane foams made from pure polyol and from pure/ regenerated polyol mixture

	Pure polyol	Pure polyol with 25 % recyclate
Tensile strength (DIN 53504), N/mm^2	4.1	5.2
Fracture straine (DIN 53504), %	441	400
Shore A hardness (DIN 53505)	62	60
Tear propagation strength (DIN 53507), N/mm^2	7.7	8
Flex life, 90 000 flexes (DIN 53543), mm	< 6	< 1
Abrasion (DIN 53516), mg	300	260

9.2.2.3. Hydrolysis and Alcoholysis of Polyesters and Polyamides

Polyesters and polyamides can be broken down to low molecular mass starting materials by hydrolysis and alcoholysis processes analogous to those used for polyurethanes [288]. The hydrolysis of polyesters yields carboxylic acids and alcohols; these can be fed to the production process only after extensive treatment operations. The hydrolysis of polyamides generates the respective amino carboxylic acids or dicarboxylic acids and diamines.

The alcoholysis of polyesters leads to the corresponding diols and dicarboxylates (Scheme 3).

If poly(ethylene terephthalate), PETP, is alcoholyzed with methanol, the monomers required for the polymerization of this plastic can be regenerated. The use of polyfunctional alcohols leads to polyols, which can be modified as required for a variety of purposes.

Processes patented by DuPont illustrate several routes for regenerating the starting materials used in PETP manufacture. For example, terephthalic acid can be recovered in ca. 96 % yield by reacting PETP waste with NaOH in ethylene glycol at 363 – 423 K [296]. With regard to physical and chemical properties, PETP made from 20 % recyclate and fresh terephthalic acid showed no significant deviations from commercial products.

DuPont has also devised a process in which PETP wastes are reacted at 398 – 408 K in a solution composed of bis(2-hydroxyethyl) terephthalate (BHT, ca. three parts by weight per part of waste), ethylene glycol (20 – 30 wt %), sodium acetate (200 – 250 ppm), and water (1 – 3 %) [297]. The plastics wastes, consisting of at least 85 % ethylene terephthalate, are broken down into the monomers and esterified to BHT. The competing dimerization of the product ethylene glycol to diethylene glycol can be suppressed by adding BHT to the reaction mixture.

Michigan Technological University has developed a process in which terephthalic acid and ethylene glycol can be regenerated from PETP by NH_4OH catalyzed hydrolysis [298]. The ground plastics waste is reacted with a steam/ammonium hydroxide mixture at 477 K under a pressure of ca. 3100 kPa. The resulting ammonium terephthalate is treated with sulfuric acid at 296 K to give terephthalic acid and ammonium sulfate. The TPA can be separated and purified; the liquor goes to an ammonia stripper. The ammonia released is dissolved in water, and the solution is reused for depolymerization of PETP. Ethylene glycol accumulates in the bottom of the NH_3 stripper and is purified by distillation.

Toray Industries has described a process by which PETP wastes are depolymerized in a melt of BHT and ethylene glycol [299]. The reaction, carried out at 488 – 523 K, leads to bis(2-hydroxyethyl) terephthalate and its oligomers. The BHT feed is said to have a glycol/acid ratio of 1.3 – 2.0. Similar values are reported for the initial PETP/ethylene glycol molar ratio. The weight ratio of the acid component of BHT to recycled plastic should lie in the range 1 : 4 – 2 : 1 at the beginning of the reaction. This process allows the reaction time (τ = 30 – 90 min, depending on the feed ratios) to be shortened and the undesired reaction of glycol to diethylene glycol and other impurities to be suppressed.

In the United States, Coca-Cola is using methanolysis to decompose PETP bottles into their basic components in a process developed by Hoechst Celanese [300]. The regenerated terephthalic acid is so pure that it is approved for reuse in the production of PETP bottles. Currently, this is the only recycling process worldwide that yields a plastics product permitted for direct- contact food and beverage use. In comparison with the processes of Goodyear Tire and Rubber (glycolysis) and Eastman Chemical Products (methanolysis), it also has the advantage that colorless stock can be regenerated from used colored plastics.

A purely hydrolytic process for a wide variety of polymer wastes [PETP, poly(butylene terephthalate), nylon 6, and nylon 66] has been devised by Celanese Mexicana [301]. The plastics wastes are fed with water (12 parts to 1 part wastes) and reacted at 521 K under a pressure of 48 bar. After a residence time of 2 h in the reactor, the wastes are decomposed to their monomers. Addition of a decolorizing agent, mainly consisting of activated carbon, enhances the colorfastness of the reaction products. The hydrolysis solutions are worked up by crystallization and distillation.

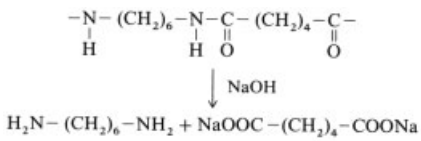

Scheme 4. Basic hydrolysis of polyamide 66 with NaOH

A process for recycling polyamides is the phosphoric acid catalyzed hydrolytic treatment of nylon 6, yielding ε-caprolactam in monomeric and oligomeric form [302], [303]. The phosphoric acid catalyst must be used in large quantities because of deactivation pro-

cesses, and thus leads to corrosion of plant equipment.

Research at BASF has shown that polyamide 66 can be recycled by basic hydrolysis [302] to give hexamethylenediamine and sodium adipate (Scheme 4). Hexamethylenediamine is directly recycled to polyamide synthesis. Adipic acid is regenerated from its sodium salt by electrolysis. The other product, caustic soda, is recycled to the cleavage reaction. These studies have progressed only to the laboratory scale and are currently being scaled up [303].

9.2.2.4. Alcoholysis of Mixed Plastics

In the preceding sections various recycling methods for sorted plastics wastes have been described. More important is the processing of mixed plastics wastes into reusable products. Such processes are of particular interest in the automotive industry because large amounts of plastics blends are generated in manufacturing and in recycling of used cars. For example, cushion and upholstery wastes consisting of a mixture of polyester, polyamide, and polyurethane arise when auto seats are manufactured. The recycling process at Ford's plant in Cologne, for example, receives 1.5 t/d of wastes. Ford, Cologne, together with the German Federal Minister of Research and Technology and the German automobile and plastics industry association, commissioned the development of a process [304] by which polymer wastes from auto seat manufacturing, partly mixed with textiles, could be processed into polyols on a demonstration scale. The polyols can be reused for cavity-filling foams in the automobile or as a feed for semirigid structural insulating foams.

This research project has led to a continuous plant with two stirred-tank reactors and a plastics capacity of ca. 100 kg/h. Figure 95 shows a flowsheet of the plant.

The weight ratio of solvent (diethylene glycol) to plastics waste was 2 : 1; in continuous operation, a maximum plastics throughput of 60 kg/h was achieved with a reaction temperature of 483 – 493 K.

The plastics are completely dissolved and partially decomposed in the first reactor. The reaction goes to completion in the second reactor. Acrylic acid was used to catalyze the alcoholysis reaction in some trials.

Both the catalytic and the noncatalytic reaction yielded a polyol that was suitable for polyurethane synthesis despite its broad molecular mass distribution.

Economic estimates based on the demonstration plant indicate that a competitively priced polyol can be produced with full utilization of a plant handling 10 – 15 t/d of plastics.

9.2.3. Conclusions and Outlook

No plant or process is currently known that offers an expedient and economical process for the hydrolysis of plastics wastes. Chemical recycling of the original monomer building blocks will not be an alternative to existing processes for some years. The obstacles are extreme reaction conditions (supercritical water) and the associated plant costs.

In contrast, two alcoholysis processes are used industrially: methanolysis of PETP and regeneration of ε- caprolactam from nylon 6 (see Section 9.2.2.). More recent research and development has shown that not only pure plastics can be recycled by alcoholysis, but also that plants can be built for the alcoholysis of plastics mixtures under mild conditions of pressure and temperature.

In the automotive industry in particular, demand for recycling processes to regenerate chemically usable feedstocks from mixed wastes is going to grow. In view of environmental problems and political pressures, it will be of interest to create plants that can handle both wastes from auto seat manufacturing (polyurethanes, polyesters, polyamides) and plastics parts from recycled cars. However, much work is needed on these processes before the marketability of the recycled products can be guaranteed.

9.3. Incineration

Incineration is the most highly developed method of treating wastes. Wastes containing organic substances can be dependably processed in such a way that the residues can be disposed of without special precautions.

Modern waste gas cleaning processes make it possible to achieve emission levels that were

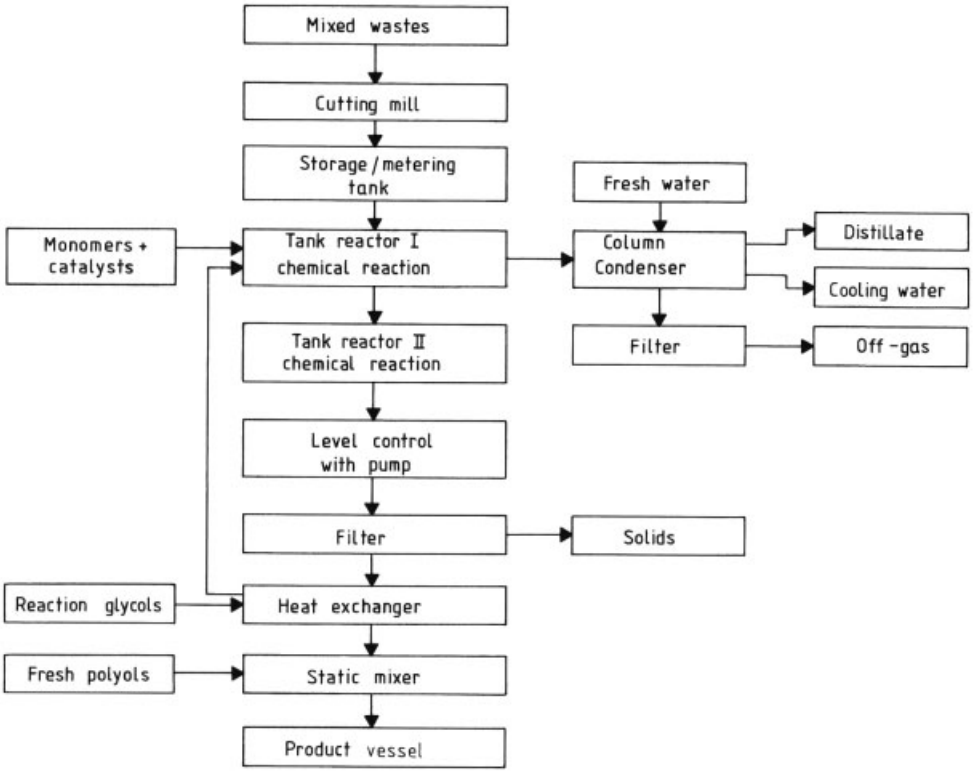

Figure 95. Block diagram of a pilot-scale plant for alcoholysis of automobile seat wastes

unimaginable just a few years ago. In industrial regions, the additional pollution due to modern waste incinerators is generally not detectable.

The following processes are employed:

1) For household and related wastes, grate firing
2) For sewage sludge, fluidized-bed firing; or this waste can be combined with household waste for grate firing
3) For industrial wastes classified as hazardous wastes, rotary kiln-firing

Industrial and commercial wastes such as packaging material, sweepings, and office and cafeteria wastes are classified together with residential wastes as "municipal-type solid wastes" or MSW and are incinerated together with household wastes in grate-fired incinerators. Small quantities of granular materials such as bleaching earths and catalysts can also be mixed with household wastes if the incineration plant is approved for this service. About 25 – 30 % of industrial and commercial wastes can be incinerated if thoroughly mixed with household waste; the heating value can be increased and made more uniform in this way.

Fluidized-bed technology is used in the large-scale treatment of sewage sludge. In Japan and Sweden, fluidized-bed furnaces are also used for MSW. Wood wastes, pretreated household waste, and contaminated soils have been incinerated in fluidized-bed pilot systems.

Solid and liquid residues from industrial production, in particular from the chemical industry, must generally be incinerated in rotary kilns if they contain organic matter and cannot be reclaimed by physicochemical treatment. This group of wastes includes oils, greases, tars, and pastelike to liquid residues from organic and pharmaceutical chemistry and from the petrochemical industry.

As a rule, centralized facilities are used; large chemical plants have incinerators close to the point where the wastes originate. In the latter case, the incinerator can be optimized for the specific properties of the residue.

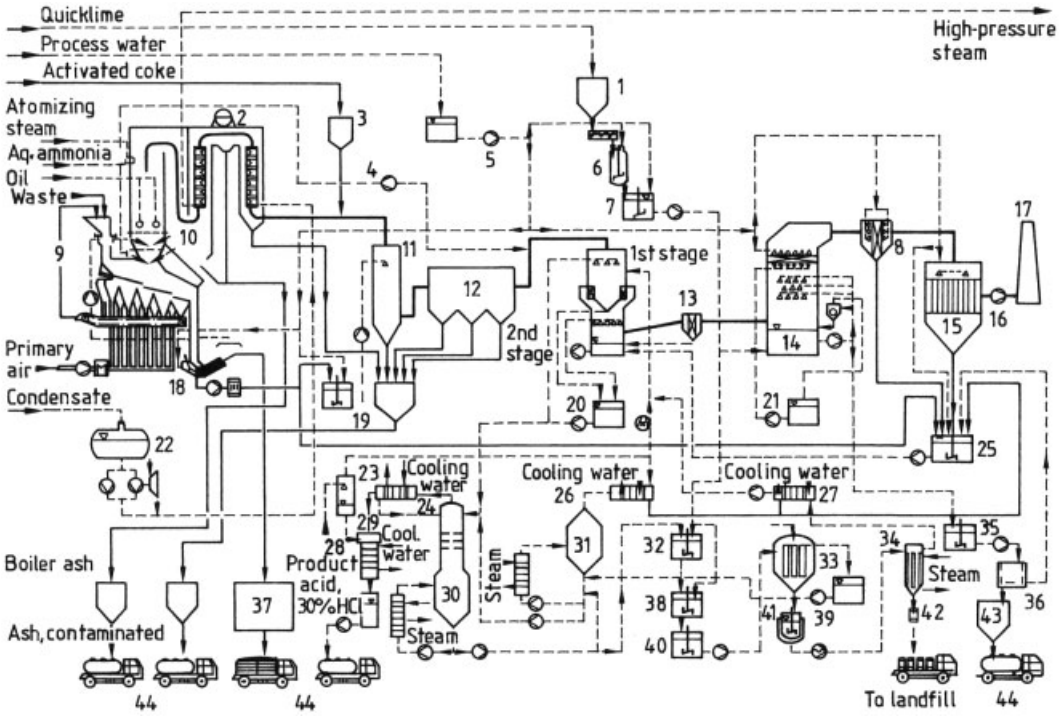

Figure 96. Flowsheet of the Hamburg-Borsigstrasse grate-fired incineration plant [308]
KEY: 1) Lime bunker; 2) Steam drum; 3) Activated coke bunker; 4) Recirculation fan; 5) Process water tank; 6) Lime slaker; 7) Milk of lime tank; 8) Mist collector; 9) Secondary air fan; 10) Boiler; 11) Evaporative cooler; 12) Fabric (bag) filters; 13) Mist collector; 14) SO_2 scrubber; 15) Wet electrostatic precipitator; 16) Forced draft fan; 17) Stack; 18) Deashing ram; 19) Supply tank; 20) External sump; 21) Tank for mist collector rinse; 22) Feedwater; 23) Scrubber; 24) Condenser; 25) Recirculating water tank; 26) Condenser; 27) Condenser; 28) Condenser; 29) Absorber; 30) Distillation column; 31) Forced- circulation evaporator; 32) Supply tank; 33) Cartridge filter; 34) Thin-film evaporator; 35) Supply tank; 36) Centrifuge; 37) Slag treatment; 38) Reactor; 39) Filtrate tank; 40) Supply tank; 41) Clarification; 42) Container; 43) Gypsum bunker; 44) Truck loading

Some wastes can serve as auxiliary fuels in industrial plants such as cement works and power stations.

The heat generated by incineration is utilized to generate power and provide remote heating.

Waste incineration has the following advantages:

1) Furnaces have been developed, built, and operated over a period of more than a century. Apparatus and methods are advanced and highly flexible.
2) Organic pollutants in the waste are destroyed with certainty by incineration.
3) Inorganic pollutants are concentrated in the residual fractions.

In high-temperature incineration processes, slag is withdrawn in liquid form, then pelletized wet. Most of the heavy metals are immobilized in this way, so that recycling or disposal of the slag presents no danger.

The melting of residues from grate-fired and fluidized-bed incinerators, which generate a mixture of ash, sinter, and slag instead of a fusible slag, is under investigation. In some cases, residues from flue-gas cleaning are added.

The process gives up to 95 % reductions in volume, so that far less landfill space is used up than if untreated wastes were disposed of.

Because waste gas cleaning methods are so advanced, emissions are below the mandated limits (in Germany those of the 17th Federal Pollution Control Regulation), often by as much as an order of magnitude.

The high level of development in waste gas cleaning means that waste incineration plants function as separation systems. The majority of the residues can be dumped in landfills. The

various residues from waste treatment are collected separately and deposited in hazardous-waste landfills if the valuable materials they contain cannot be reclaimed. In future, recovery of these materials will be required. Examples are the production of hydrochloric acid, sodium chloride, and gypsum. The metallurgical recovery of heavy metals from fly ash appears feasible in principle and is now being studied.

9.3.1. Grate Firing

Grate firing is the most widely used process for thermal treatment of wastes. Germany now has 51 plants in service; worldwide there are some 500. Figure 96 is a flowsheet of the Hamburg-Borsigstrasse plant, which came on stream in 1994 and represents the state of the art. The main steps in the process are as follows:

1) Waste receiving and preparation: scale, bin, crane, rotary cutter
2) Combustion and heat recovery: feeder, grate, air supply, combustion chamber, boiler
3) Waste gas cleaning: evaporative cooler, fabric filter, acid scrubbing, basic scrubbing, aerosol separation, SNCR denitrification
4) Wastewater treatment: HCl and gypsum recovery, wastewater evaporation

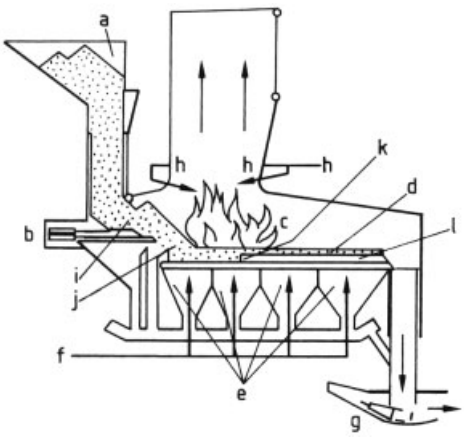

Figure 97. Grate-fired incinerator [308]
a) Waste feed bin; b) Ram feeder; c) Combustion chamber; d) Grate; e) Underblast zone; f) Primary air system; g) Slag discharge; h) Secondary/tertiary air; i) Refuse bed; j) Drying/ignition zone; k) Main combustion zone; l) Burnout zone

Combustion and Heat Recovery. The most important component of a grate-fired incinerator is the grate, which conveys the waste in a horizontal or inclined direction after the waste has been removed from the bin by a crane and has passed through the feed hopper and feeder. The grate is carried on a supporting structure built into the combustion chamber side walls in the underblast region; the grate proper consists of individual grate bars or grate rolls (Fig. 97).

Grate incinerators are classified according to how the charge is conveyed and rabbled. Types include reciprocating-bar stokers, single-stage and multistage, forward and reverse feed; pivoting step grates; single-stage and multistage traveling grates; roller grates; and reciprocating recirculating grates.

The grate itself can be horizontal, inclined down toward the end, or inclined up toward the end in the combustion chamber. It is bounded in front by the refuse feeding device (charging system), in back by the discharge device, and at the sides by walls, which can be uncooled, water-cooled, or air- cooled. The ash catchers are located below the grate; these also bound the underblast zones. Primary combustion air is delivered to the ash catchers through control valves; it is distributed beneath the grate as appropriate to the combustion process and forced by pressure into the combustion chamber through openings in the grate (free grate area).

Primary air passes through the grate in cross or oblique flow. The waste undergoes several partly overlapping processes: drying, degassing, gasification, combustion, and finally sintering to slag. These parallel reaction processes interact indirectly via the combustion chamber space above the grate, so that they can be well, and to a great extent individually, controlled even though there are no sharp boundaries between them.

Grate firing has the disadvantage that the generation of reaction products is largely uncontrollable. In the boundary zone between the drying and combustion steps, degasification processes generate carbon compounds that cannot be completely destroyed in the combustion space. A similar situation occurs in the boundary zone between combustion and cooling, where the production of sinter is a more or less random process that depends on process control and the quality of the feed material.

In modern incinerators, primary air is admitted substoichiometrically. The unburned gases generated as a result, chiefly CO and H_2, are combusted by admitting secondary air to the combustion chamber. This two-stage combustion system gained acceptance in the 1960s. The reason was that the heating value of waste increased so that the temperature in the charge bed often exceeded the melting point of the slag, which formed deposits on the grate and the furnace side walls, interfering with operation.

Grate-fired incinerators must meet the following requirements:

1) The primary air flow rate must be variable over sufficiently wide ranges with respect to the length and width of the bed of material on the grate
2) The primary air supply must not be impaired even after prolonged mechanical and thermal stress resulting from normal operation
3) The outlet holes for primary air in the grate structure must not become plugged or enlarged during operation
4) The rate of bed movement on the grate must be separately variable for each grate zone
5) The depth of the waste bed on the grate and the position of the combustion zone must be controllable
6) The waste should be thoroughly stoked and turned
7) The combustion process must generate as little fine dust as possible
8) There must be provision for keeping the waste/slag layering uniform over the entire area of each grate zone
9) Primary air channeling must be avoided
10) The quantity of material falling through the grate must be small and there should be a reliable means for removing this material
11) The primary air should adequately cool the grate under all operating conditions
12) The system must also have the capability of running on preheated combustion air and with recirculating off-gases admitted at selected points
13) Control should be largely automated and the system must respond immediately to control actions
14) Rapid startup and shutdown of the incinerator must be possible
15) Sintering and deposition of material on the grate, as well as scaling and corrosion of the grate, must be prevented
16) The grate bars must be easily and quickly replaceable
17) Reactor runs should be as long as possible

Combustion Chamber. Combustion takes place in the combustion chamber, which is bounded below by the grate, to the sides by cooled or uncooled walls, and above by a roof or boiler heating surfaces. The heat evolved here must be removed.

Only a small amount of waste is burned on and immediately above the grate, since the content of volatiles is high. The main process in the incinerator is combustion of volatiles in the combustion space above the solid waste. External, visible signs of this type of combustion are long flames in the combustion space, often up to a meter high. In contrast, the flame length when bituminous coal is burned is just a few decimeters, depending on the quality of the fuel (i.e., the content of volatiles).

The gas flow pattern is characterized as cocurrent, countercurrent, or mixed (central feed), depending on the geometry of the combustion space, which is defined by the following parameters:

1) Shape and size of the grate. The combustion space is bounded below by the grate, whose length and inclination largely determine its geometry.
2) Process zones on the grate. To allow for different-sized drying, degassing, combustion, and burnout zones, the combustion space must permit fast introduction or removal of heat.
3) Turbulence and homogenization of flue gas flow. Thorough mixing of the off-gases is essential for their complete combustion.
4) Residence time of combustion gases in the hot combustion zone. An adequate reaction time at high temperatures must be provided to ensure complete combustion of gases and solids.
5) Partial cooling of off-gases. To prevent sintering of fly ash to the walls and in the boiler, the off-gas temperature at the end of the combustion space must be limited.

The design of the combustion chamber is a compromise because the requirements change with the properties of the fuel. The gas flow pattern (cocurrent, countercurrent, mixed) is chosen to suit the actual requirements.

Because of the risk of deposits and corrosion, there are no internals in the combustion space or the first boiler flue pass. Heat transfer to the boiler in these zones is radiative.

Because the off-gases are deflected several times before entering the convection section, the gases are thoroughly mixed and the temperature is distributed over the cross section. Because the volume of the combustion space plus first, second, and third flue passes is very large, the residence time of the combustion gases in the radiant section is prolonged while the thermal loads per unit volume, unit cross section, and unit heating surface are reduced. The large gas-side flow cross sections permit low gas velocities, resulting in a low dust load in the gases downstream of the combustion space.

In continuous operation with contaminated heating surfaces, the off-gas temperature upstream of the convection heating surfaces is ca. 650 °C. The transverse and longitudinal pitch of the tubes in the convection section is large in comparison to the tube diameter. The horizontal arrangement of the convection section also promotes the settling of fly ash, with a beneficial effect on the downstream gas- cleaning system. Fly ash is collected in fly-ash hoppers beneath the heating surfaces and discharged.

Combustion Air Supply. Only a small amount of combustion takes place on the grate itself. The volatile constituents of the fuel escape from the bed of refuse when it is heated and burn on their way through the combustion space. This combustion cannot be controlled when underblast is used. The air flow pattern also governs whether there is a risk of corrosion in the combustion space cooled with boiler tubes. Hydrochloric acid generated by the burning of wastes can attack the tubes if incompletely burned gases (carbon monoxide, CO) are simultaneously present. The supply of combustion air must therefore be sufficient to promote complete combustion of these gases. Grate-fired incinerators are always operated with excess air ($n = 1.5 - 2.0$).

The combustion air must perform the following tasks:

1) Serve as oxidizing agent
2) Cool grate and side walls
3) Prevent slagging in combustion chamber
4) Mix the off-gases and prevent bypassing

A variety of air and air-distribution systems make it possible to achieve these goals.

Primary air is commonly drawn from the refuse bunker. The bunker doors are usually closed, so that a vacuum is set up in the bin region, reducing odors. A filter screen in the intake duct removes coarse entrained particles. A blower delivers the primary air into the underblast zones through the pressure duct. Air preheating is employed when the heating value of the waste drops to the point that predrying is required. Primary air is injected through the grate into and through the bed of fuel, cooling the grate bars and promoting gasification and combustion in the bed.

Secondary air is injected into the combustion space through nozzles to ensure complete combustion of the off-gases, provide intensive mixing of combustion gases, and thus prevent bypassing of unburned gases.

A second stream of air, also drawn from the bunker, is blown onto the side walls of the combustion chamber to keep the flame away from the walls, cool them, and promote the oxidation of the off-gases in the combustion space. If the furnace walls are perforated, the sidewall air can be injected into the combustion chamber perpendicular to the wall over a large area, thus inducing turbulence in the combustion gases so as to break up bypassing jets of reducing gases.

Fly ash is kept away from the walls in this way, preventing formation of deposits (fly ash becomes consolidated if it touches the side walls). Simply cooling the wall, for example by extending the boiler structure downward, does not have the same effect but does lower the wall temperature.

Combustion Control. Waste incineration consists of successive subprocesses. Combustion and steam generation are the main processes and are closely interlinked. The combustion rate and the rate of energy release must be held as constant as possible for uniform steam produc-

tion. The specified steam parameters must be maintained to ensure trouble-free operation.

Boiler and Heat Recovery. Incinerators for household and hazardous wastes can be fitted with steam generators or hot-water boilers. The types of steam boilers used are natural- circulation, forced- circulation, and once-through. Problems with the water recycle stream due to the fluctuating heat delivery have not arisen. The forced- circulation type of boiler used initially has been replaced by the simpler natural- circulation type. Thermal loads per unit heating area are relatively low because of the large volumes of refuse incinerator boilers, so that circulation in the natural- circulation boiler is adequate even with severe load variations.

Older trash incinerators were provided with waste-heat boilers. The refuse was burned in uncooled furnaces and the combustion gases were cooled in downstream heat-recovery units. With increasing heating value of the waste, combustion temperatures have risen to the point that uncooled linings cannot withstand the thermal loads. Cooling was needed for the combustion chamber walls, and so the boiler was placed directly over the grate.

The off-gases must have a residence time of at least 2 s at 850 °C so that all organic species, particularly dioxins and furans, can be completely destroyed.

Turbulence is induced in the off-gases to provide backmixing and also prevent the formation of a reducing atmosphere (CO bypassing) that would promote boiler tube corrosion.

Because of the fluctuating rate of heat evolution, turbulence induced in the off-gases by secondary air injection or deflecting structures is not sufficient by itself. The use of a large excess of air prevents air starvation under all loading conditions that occur. In addition, a rammed lining is present in the lower portion of the cooled combustion space to protect against corrosive attack [305].

A special feature of waste incinerator and boiler installations is the high dust content of the off-gases, since for process engineering reasons the slag softening point must not be exceeded. Precautions are therefore needed to prevent flue-gas contamination.

The combustion gas velocity in the first boiler flue pass is low so that fly ash can settle by gravity. The gas temperature upstream of the convection section must be less than 750 °C, because at higher temperatures the entrained contaminants remain doughy in consistency and tend to build up into deposits that cannot be removed by the cleaning equipment.

For wastes with higher heating values, the combustion gases do not cool to this level in the combustion chamber. An idle pass is therefore inserted in which the gases are completely combusted and homogenized. Fly ash is separated in the flow-reversing zones.

The off-gas velocity must not exceed 10 m/s, since the heating surfaces become fouled at higher velocities. Boilers in waste combustion plants can also be equipped with horizontal passes.

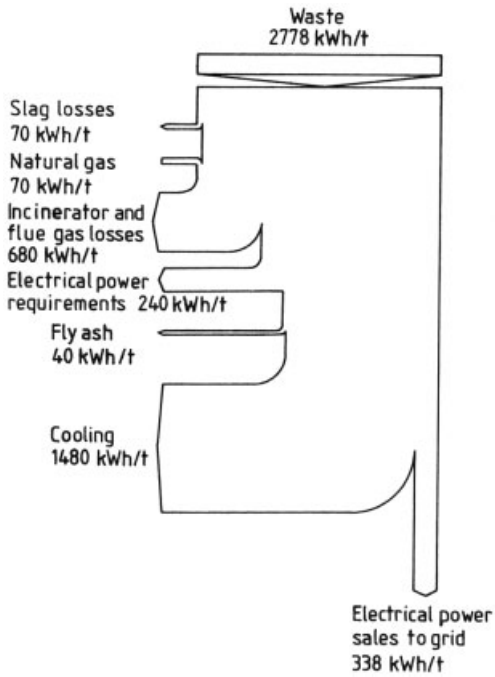

Figure 98. Energy balance for a grate-fired incinerator with flue-dust melting [309]

The boiler must be designed for high temperatures and pressures in order to utilize the energy contained in the waste. Systems have been built for up to 78 bar and 500 °C. These extreme values, however, lead to serious corrosion problems, so that 42 bar and 450 °C should not be exceeded without special justification [305].

The steam generated can be used to generate electricity and/or for district heating. The best efficiency is achieved in a cogeneration arrangement with a pass-out condensing turbine; Figure 98 shows the energy balance for this scheme, including natural gas used to fuse the fly ash.

Off-Gas Cleaning. The thermal treatment of wastes is a complex chemical process. The main reaction, combustion, is an oxidation with the typical combustion products carbon monoxide, carbon dioxide, water vapor, sulfur dioxide, and nitrogen oxides. Along with these compounds, other pollutants are produced; depending on the composition of the waste, these include hydrogen chloride, hydrogen fluoride, dust, heavy metals, nitrogen oxides, dioxins, and furans.

Even though dust-collection and off-gas cleaning technology has reached a high level, further steady progress has been seen since the 17th Federal Pollution Control Regulation (17. BImSchV, see → Air, Chap. 4.3.9.) took effect. The limits set by this regulation led to the addition of denitrification (NO_x removal) and dioxin abatement to treatment systems, and emission levels are now considerably lower than the mandated limits. The recovery of valuable residues is also a goal.

Off-gases are cleaned through a combination of various steps:

1) Dust collection in electrostatic filters or baghouses
2) Wet absorption of acidic species (SO_2, HCl, HF) in multiple scrubbing stages varying in pH
3) Dry sorption of trace organics (such as PCDD/PCDF), heavy metals, and acidic species on activated carbon filters in moving-bed adsorbers and entrained-bed adsorbers combined with bag filters
4) Catalytic reduction of nitrogen oxides (SCR)
5) Noncatalytic reduction of nitrogen oxides by oxidation in the combustion chamber (SNCR)
6) Catalytic oxidation of trace organics such as PCDD/PCDF (OxKat)
7) Thermal treatment (incineration) of pollutant-laden sorbents (residues)
8) Thermal treatment of residues (e.g., pollutant-laden sorbents) from off-gas cleaning
9) Thermal treatment (melting) of slags and dusts
10) Recovery of HCl, NaCl, and gypsum from wet scrubber residues

The requirement profiles for these processes are stated and commercially available equipment is described in [308].

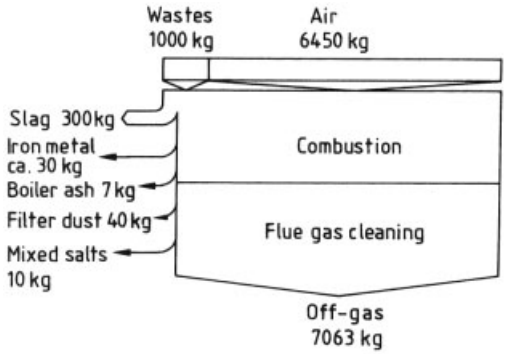

Figure 99. Mass balance for a grate-fired incinerator [309]

Residues from Grate Firing. Combustion of wastes on the grate yields solid, liquid, and gaseous reaction products. Figure 99 presents the mass balance. The literature does not use consistent terminology for solid residues. The following terms are employed here [308]:

1) Slag: solid residue left after waste combustion and not entrained in flue-gases. It is discharged from the end of the grate or falls through the bars between grate steps. This term is not entirely correct since it refers to a mixture of ash, sinter, and slag.
2) Boiler ash: solid residue formed on boiler heating surfaces and falling into hoppers beneath the flue passes.
3) Filter dust: solid residue obtained by dust collection from raw gases and discharged via the hoppers beneath the electrostatic or bag filters.
4) Fly ash: sum of boiler ash and filter dust from electrostatic or bag filters before chemical flue-gas cleanup. The terms "filter dust" and "filter ash" are used as synonyms for fly ash in the literature; boiler ash makes up only a small fraction of the dust (ca. 10 – 20 wt %)
5) Reaction product: solid residue from dry or quasi-dry stack-gas cleanup with upstream dust collection. This product is generated in

the dry or quasi-dry reactor and in downstream dust collection equipment. Concentrated residue from wet stack-gas scrubbing is also referred to as reaction product if the scrub liquor is concentrated in a spray dryer. This reaction product is obtained in the dryer and in the dust collection equipment immediately downstream of it.

The solid residue from dry or quasi-dry flue-gas cleaning without upstream dust collection should therefore be termed a mixture of flue dust and reaction product.

Many publications refer to this mixture as reaction product or dry/quasi-dry residue; however, these expressions do not indicate whether the product contains flue dust.

The quantity and composition of combustion products depend on the makeup of the fuel, the mode of combustion, and the type of stack-gas cleanup.

Table 25. Composition of solid residues from grate firing [309]

Element	Slag	Fly ash
Silicon, wt %	1 – 32	10.28 –16.82
Aluminum, wt %	0.46 – 18.00	6.39 – 9.5
Calcium, wt %	3.32 – 13.4	2.7 – 11.7
Sodium, wt %	0.5 – 6.0	1.2 – 7.5
Magnesium, wt %	0.39 – 1.82	1.2 – 3
Potassium, wt %	0.29 – 2.14	1.9 – 7.8
Iron, wt %	3.71 – 23	2.5 – 4.6
Copper, wt %	0.04 – 0.7	0.07 – 0.2
Chromium, wt %	0.015 – 1	0.02 – 0.32
Nickel, mg/kg	39 – 760	50 – 460
Zinc, wt %	0.05 – 2.1	0.62 – 5.5
Lead, wt %	0.03 – 1.7	0.24 – 5.4
Tin, mg/kg	200 – 1 700	500 – 6 200
Cadmium, mg/kg	0.1 – 79	60 – 1 000
Mercury, mg/kg	0.07 – 7 000	0.2 – 20
Chlorine, wt %	0.03 – 0.63	0.6 – 8.2
Phosphorus, wt %	0.35 – 3.37	0.02 – 3
Sulfur, wt %	0.2 – 0.41	0.6 – 2.97
Fluorine, mg/kg	20 – 1 100	30 – 12 000
Carbon, wt %	0.5 – 5	1.6 – 4.2

Table 25 presents the elemental compositions of solid residues; most of the elements are present as oxides. This form of analysis was chosen because the type of compound (oxide, chloride, sulfate, etc.) cannot be exactly stated for each element and, moreover, cannot be determined by element-specific analytical methods such as atomic absorption spectrometry and ion chromatography. If other elements totaling less than 1 wt % are neglected, the balance consists of oxygen, since all relevant nonoxide compounds (chloride, fluoride, phosphate, sulfate, sulfite, etc.) have already been considered.

Reaction products from dry and quasi-dry flue-gas cleaning are declining in importance because the next generation of refuse incinerators will employ wet scrubbing, so that hydrogen chloride and sulfur dioxide can be separated in different stages and the scrub liquors can be converted to products such as hydrochloric acid, table salt, and gypsum.

9.3.2. Fluidized-Bed Firing [306]

Since it was devised by Fritz Winkler over 70 years ago, fluid-bed technology has found wide use in chemical processing, drying and cooling, gasification, and combustion. It is a proven technique for burning low-grade coals and sewage sludges. More recently it has also come into use for the incineration of household wastes, light fractions of household waste (Fig. 100), and contaminated wood and soil [309].

The fluidized-bed furnace is a lined combustion chamber in the form of a vertical cylinder. In the lower section, a sand bed on a grate or distribution plate is fluidized with air. The waste for incineration is continuously fed into the fluidized sand bed from the top or side.

If necessary, secondary air can be injected into the reaction space above the fluid bed to ensure complete combustion. Oversize material drops out of the lower part of the reactor; ash-laden flue gas exits from the upper part.

Primary air admitted through the perforated bottom thus fluidizes the bed and supports combustion. If the air flow rate in the bottom portion of the bed is substoichiometric, the material for combustion first experiences partial or complete degassing. The waste is then completely burned when secondary air is admitted in the upper zone of the reactor.

As in the grate-fired case, the quality and quantity of the products depend on the nature of the feed, the reactor itself, and the operating conditions prevailing therein.

The input to the process is characterized by

1) The physical properties of the waste and inerts:
   a) Particle size
   b) Particle shape

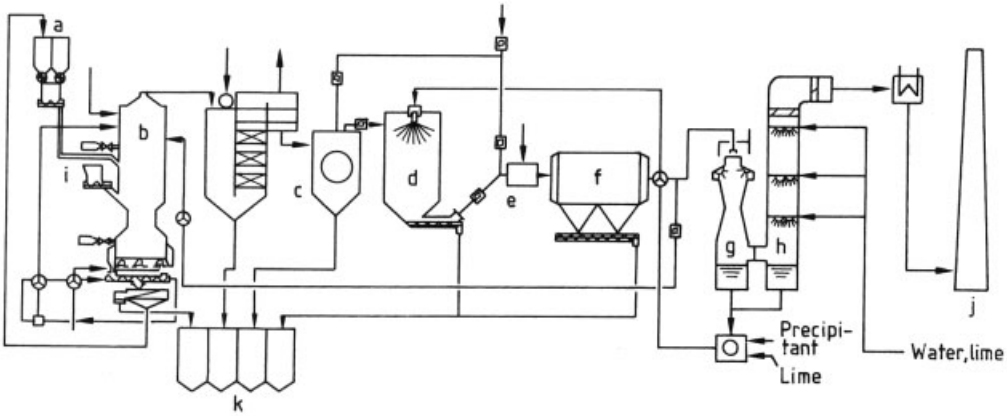

Figure 100. Flowsheet of the fluidized-bed incineration plant for prepared residential and municipal-type wastes at Berlin-Ruhleben [309]
a) Sand and dolomite bunkers; b) Rotating fluidized-bed incinerator; c) Cyclone; d) Spray dryer; e) Entrained-bed reactor; f) Fabric filters; g) Venturi scrubber; h) SO_2 scrubber; i) Refuse feed; j) Stack; k) Waste bunkers

c) Bulk density
d) Heating value
e) Chemical composition

2) The state parameters of the combustion air:
   a) Density
   b) Viscosity
   c) Temperature
   d) Flow velocity

The reactor is characterized by:

1) Dimensions
2) Shape
3) Air supply type

The operating parameters are the following:

1) Pressure drop
2) Limiting velocities at the fluidization point and the discharge point
3) Depth of the fluid bed
4) Temperature distribution

Fluidized-bed systems have the following features:

1) Complete mixing of solid particles among themselves and with combustion air
2) Resulting good mass and heat transfer between fuel particles and between fuel and combustion air
3) Fixed reactor with no moving parts in the combustion space
4) Possibility of using low-grade fuels with high inert contents
5) Fast process control via throughput
6) Lower air excess than in grate-fired systems; correspondingly small flue-gas streams and low pollutant burdens
7) Possibility of heat removal in bed zone
8) High ratio of hot material to inlet fuel in fluid bed
9) Good ignition due to turbulent motion of fuel; fast combustion due to uniform temperature in fluidized bed
10) Possibility of using inert materials to trap pollutants in the furnace

Because of the special conditions in the fluidized bed, some technical measures differ from those in grate firing:

1) The fuel particle-size distribution must be kept narrow
2) Problems may arise in large-cross-section reactors if the inlet fuel is not adequately distributed over the cross section
3) Solids can be carried out of the bed in an uncontrolled manner (entrainment)
4) The vigorous motion in the combustion chamber can cause erosion
5) Erosion can occur in downstream equipment because of the high dust load in the stack gas
6) The combustion temperature must be held below the sintering point of the fuel so that agglomerates do not form and collapse the bed

For start-up, the cold fluidized bed is heated with gas- or oil-fired auxiliary burners. Once the operating temperature has been reached, the feed material is introduced, and is dried, degassed, and burned in the bed. As combustion proceeds, the particle size decreases until the ash particles are entrained by the upward gas flow and carried out of the bed. Constituents that do not burn and cannot be lifted sink to the bottom of the reactor where they can be discharged.

In combustion operation, the fluidized bed consists of ca. 1 – 3 vol % unburned fuel and ca. 97 – 99 vol % ash. If the amount of ash is sufficient to form a fluid bed, inorganic substances need not be used as bed material.

Ash and oversize can, however, build up even in a fluid-bed incinerator. An excess of these materials can impair the bed, so they are discharged from the lower zone of the fluidized bed.

In comparison with grate-fired incinerators, fluidized-bed units exhibit a higher dust burden in the hot flue gas at the exit from the reactor. This must be considered in designing the flue-gas cleaning equipment.

The discharge behavior is controlled by means of the inlet air velocity, but the discharge of fines with the reaction gases cannot be prevented. The dust burden of this raw gas may be of the order of 20 – 80 g/m^3. Stationary or slow fluidized-bed firing is performed with gas velocities of up to ca. 2.5 m/s, which give a high-density bed with a well-defined surface.

The heat of combustion can be utilized by recovering the sensible heat of the flue gases in a downstream boiler and through heating surfaces in the walls. (The walls can contain membrane tubes or finned tubes as part of the steam-generator circuit). The combustion air in such a unit is split into two streams; primary air fluidizing the bed and secondary air injected into the space above the bed to support complete combustion.

Fluidized-Bed Firing Systems. Fluidized beds are classified as follows:

1) "Stationary" (bubbling or turbulent) fluidized beds (Special case: rotating fluidized beds in single- and two-bed versions)
2) "Fast" fluidized beds
   Circulating beds
   Expanding beds

Figure 101 illustrates the principles on which these firing systems operate.

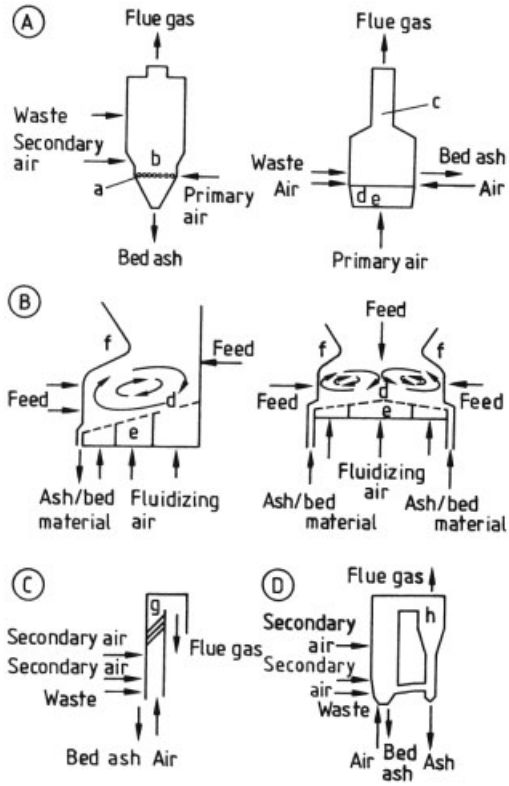

Figure 101. Fluidized-bed incinerator types
A) Stationary fluidized bed; B) Rotating fluidized bed (special case of stationary fluidized bed); C) Expanding fluidized bed; D) Circulating solids fluidized bed
a) Distributor pipes; b) Ash hopper; c) First boiler pass; d) Distributor plate; e) Air chambers; f) Deflector plate; g) U-beam particle collector; h) Hot cyclone

The stationary fluidized-bed reactor of conventional design is the basic form in fluidized-bed technology. It consists of a cylindrical or rectangular reactor bounded at its lower end by a distribution plate.

Stationary-bed units are less sensitive to feed particle size than other fluidized-bed types; they can accordingly be used for a wide spectrum of particle sizes, but household waste must at least be ground before incineration.

The rotating bed is a special case of the stationary bed in which forced circulation (rotation) improves solids mixing and transport.

Rotating fluidized-bed reactors are made with one or two beds. The combustion chamber is rectangular in cross section. One or two side

walls of the reactor bend inward above the inclined distribution plate to form deflectors parallel to the distributor, thus defining the position of the fluidized bed.

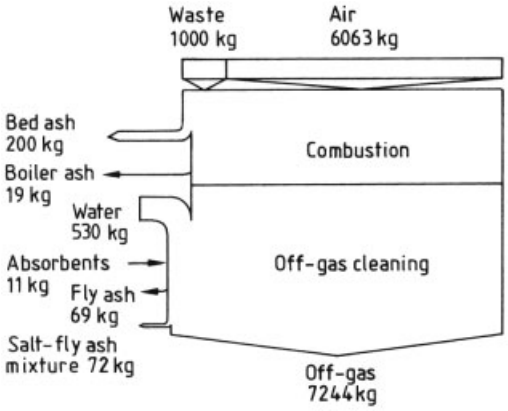

Figure 102. Mass balance for a fluidized-bed incinerator [309]

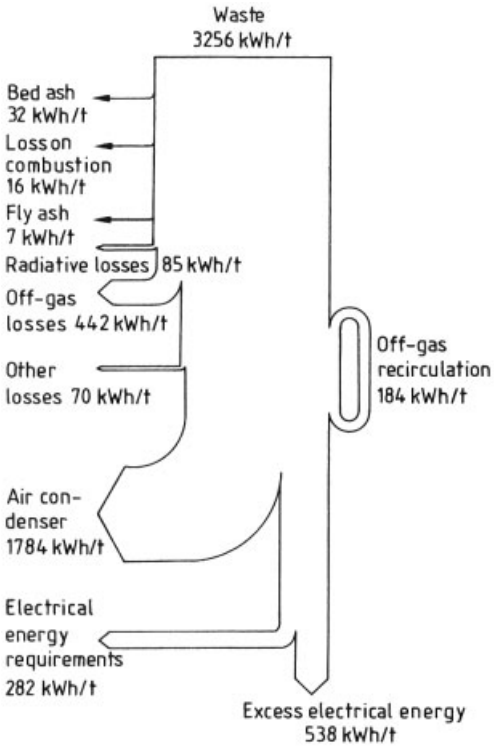

Figure 103. Energy balance for a fluidized-bed incinerator [309]

Primary air is admitted to the reactor through several air chambers; the flow velocity increases toward the lower end of the distribution plate, so that the bed material there is lifted up against the deflector and made to move toward the higher end of the grate, where the flowvelocity is lower. In this way, the bed is caused to move in an elliptical orbit. In the single-bed variant, the fluid bed thus comprises a rotating "roller"; in the two-bed version, it takes the form of two counterrotating rollers. Bed rotation can be aided by injecting secondary air under the deflector plate or opposite to the restriction.

In contrast to conventional stationary fluidized-bed firing, this design has an additional horizontal fluidizing component. The waste can thus be distributed more evenly and the bed temperature equalized. Components that are not lifted or burned move downward along the inclined distribution plate and exit through an ash discharge at the bottom end of the plate. Ash is discharged continuously or intermittently through a water - cooled screw feeder and classified; the fine fraction is recycled to the reactor.

The bed temperature is ca. 850 °C. For a given heating value, the temperature is controlled by recirculating the stack gases to the primary air system so as to limit the heat generated in the bed. The availability of oxygen is diminished in this way and combustion becomes a two-stage process:

1) Drying, degassing, gasification, and partial burning of fuel in the fluidized bed with primary air
2) Complete combustion in the zone above the bed with secondary and tertiary air

The properties of the rotating fluidized-bed incinerator can be summarized as follows:

1) Rotation of the bed effects good transverse distribution, so that only one feed location is needed
2) High heat-transfer coefficients with large heat-transfer surface areas result in fast heating of the feed
3) The low bed temperature (ca. 800 °C) improves the trapping of acidic pollutants such as SO_2, HCl, and HF in the bed
4) The bed has a uniform temperature distribution because of avoidance of unequally hot spots and avoidance of spotty ignition, smol-

dering (reducing atmosphere), and sintering effects

5) The high degree of turbulence and high velocity of the flue gases in the throat above the bed (flow velocity increases from 2.6 to 5.4 m/s in the throat, then drops to 1.96 m/s in the disengagement space) leads to:
   a) Longer residence time
   b) Complete burnout
   c) Temperatures not exceeding the ash softening point, resulting in absence of sintering effects and no fouling of boiler heating surfaces
6) Splitting the combustion air into primary and secondary air and recirculation of cooled stack gases yields stepwise combustion and thus:
   a) Good controllability
   b) High flexibility
   c) Low NO_x levels
   d) Tolerance to variations in fuel heating value, moisture content, and ash
7) There are no moving parts in the high-temperature zone
8) The design offers high reliability and availability
9) The removal of nonflammable constituents is promoted

Emissions from Fluidized-Bed Reactors. When wastes are incinerated in the fluidized bed, pollutants are emitted in the flue gas, with the solids path, and in the wastewater. Emissions depend on the pollutants introduced with the fuel, the way in which fuel is inlet, the geometry of the combustion space, the reaction properties, the control response of the reactor, and the method used for flue-gas cleaning. Figures 102 and 103 show the mass and energy balances.

The pollutant content of the fuel has the most marked effect on emissions, but the inhomogeneous makeup of wastes means that this effect can only be quantified case by case. The preparation of the feed material for fluidized-bed incineration consists, at least, of grinding to the requisite particle size. This operation makes the combustion behavior more uniform and has a beneficial effect on emissions. Other preparation, such as controlled removal of pollutants, reduces emissions.

The parameters of the bed process also influence emissions such as carbon monoxide, organics, and in part NO_x and solids (fly ash and soot). Stepped admission of combustion air over the reactor height and distributor plate cross section offer some opportunities to optimize the combustion process, for example to achieve complete burning with even a small air excess, to minimize the volume of reaction gases, and to cut down the amount of flue dust discharged.

The carbon monoxide level in the stack gas serves as an indicator of incomplete combustion. The content of CO in the product gas from fluidized-bed incineration can be lowered without increasing the air/fuel ratio if secondary air is admitted in stages.

Sulfur oxides can be trapped in the bed by adding limestone. The binding behavior depends on several parameters, such as the temperature in the fluidized bed and the Ca/S ratio. Widely varying efficiencies have been reported.

The majority of the nitrogen is introduced with the combustion air and reacts with atmospheric oxygen at 1200 °C and higher to form nitrogen oxides. Because incineration temperatures generally lie between 750 and 900 °C, with a maximum of 1200 °C, this reaction does not have to be considered. The NO_x content of the flue gas is therefore due to the largely temperature-independent reaction of fuel nitrogen with atmospheric oxygen. Nitrogen oxide formation increases when limestone is added directly to the combustion space to trap sulfur.

The halogens chlorine and fluorine present in the fuel are almost completely liberated at the prevailing combustion temperatures; they can be partly removed if limestone is added directly to the bed. The optimum temperature range for chlorine and fluorine trapping is ca. 300 °C. Because combustion temperatures in the bed are higher, the expected capture rates for these elements are low; estimates are 50 – 80 % for chlorine and 90 – 99 % for fluorine.

The heavy metals stream is split between flue gases and bed ash. Fluid-bed incineration yields a higher fly-ash content than grate firing, so more heavy metals must be anticipated in the flue dust (depending on the temperature); see Table 26. Fly ash from fluidized-bed combustion must therefore be completely collected.

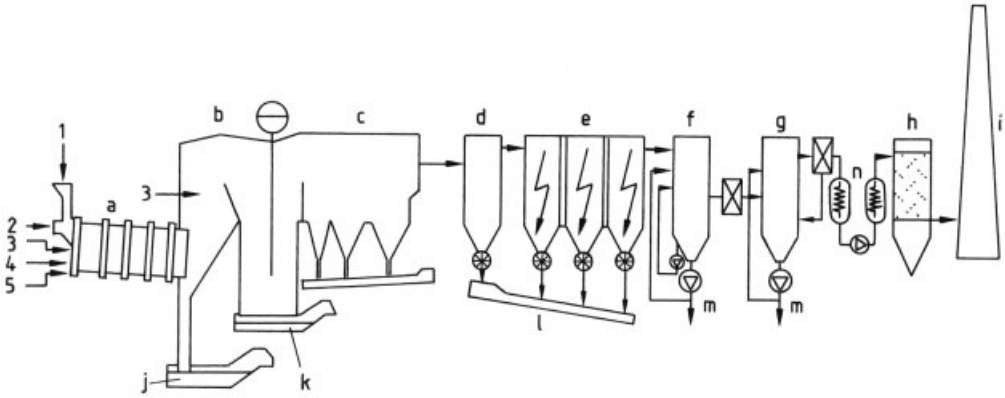

Figure 104. Flowsheet of a rotary-furnace system for hazardous-waste incineration [308]
a) Rotary kiln; b) Post-combustion chamber; c) Waste-heat boiler; d) Quenching; e) Electrostatic precipitators; f) Wet scrubber (acid); g) Wet scrubber (basic); h) Activated-carbon filter; i) Stack; j) Slag discharge (wet); k) Ash discharge (wet); l) Ash discharge (dry); m) Mist collector; n) Heat exchangers
1) Solids; 2) Drums; 3) High-heating-value liquid wastes; 4) Pastelike wastes; 5) Low-heating-value liquid wastes

Table 26. Composition of solid residues from fluidized-bed firing (in wt %) [309]

Element	Bed ash	Fly ash
Silicon	47 – 70	
Aluminum	0.14 – 21	
Calcium	5.5 – 13.1	
Sodium	0.8 – 4.0	
Magnesium	0.1 – 2.4	
Iron	0.20 – 0.65	
Zinc	0.0027 – 0.201	
Lead	0.0055 – 0.0812	0.066 – 0.088
Cadmium	1.85 – 5	13.0 – 15.5
Chlorine	0.01 – 0.25	0.3 – 0.6
Phosphorus	0.3 – 1.2	
Sulfur	0.1 – 0.7	0.4 – 0.8
Carbon	0.06 – 0.09	1.6 – 1.8

Emissions by the solids path are divided among these streams (Fig. 102):

1) Bed ash
2) Fly ash
3) Residues from pretreatment

Fly ash entrained in the flue gases is finer-grained than bed ash. Its content of pollutants depends on the time at which the stack gases are cooled. If fly ash is collected prior to flue-gas cooling, it has a pollutant content comparable to that of bed ash. If the flue gases are cooled in the boiler, volatile heavy metals begin to condense here, and under some circumstances new organic compounds begin to form as well; these are predominantly adsorbed on the surfaces of fly-ash particles. The formation of dioxins and furans after cooling of the stack gass (de novo synthesis) is especially relevant. The composition of filter dusts produced in stack-gas cleanup depends on the cleanup method among other factors.

Off-gas cleaning is comparable to the process with grate firing (see Section 9.3.1).

9.3.3. Rotary-Kiln Firing

A rotary kiln consists of a cylinder inclined in the transport direction and lined on the inside with refractory material or a cooled steel shell. The interior of the cylinder can be up to 20 % filled with fuel, which forms a moving bed. Rotation of the furnace about its longitudinal axis turns over the contents and, by virtue of the inclination, causes it to move toward the lower end.

Rotary kilns have been used for around 50 years in the cement, lime, gypsum, magnesite, dolomite, and iron industries. More than 1000 such furnaces are in service.

This type of equipment can be used for incinerating all kinds of wastes: solid, pastelike, and liquid. The furnace is insensitive to variations in the consistency and composition of mixed wastes.

Rotary kilns are especially useful when incineration plants are being designed for future wastes whose composition is not known but must be allowed for, as is particularly the case

with hazardous wastes. Figure 104 is a schematic flowsheet of such an installation. Throughput varies as a function of inlet material, furnace design, and operating conditions.

A rotary-kiln plant (Fig. 105) consists of:

1) Feed devices
2) Air supply and sealing devices
3) Auxiliary firing
4) The rotary shell and liner
5) The drive
6) The afterburner
7) The ash and slag discharge devices

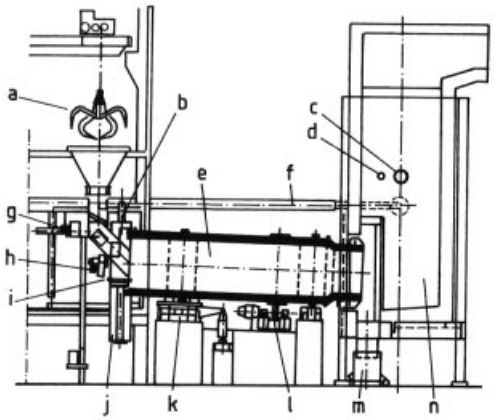

Figure 105. Schematic drawing of an AE & E hazardous waste incinerator [308]
a) Charging by crane; b) Primary air system; c) Oil burner; d) Wastewater injection; e) Rotary kiln; f) Secondary air system; g) Drum charging; h) Oil burner; i) Injection of solvent sludge, pastelike wastes; j) Head wall of rotary furnace; k) Thrust rolls and longitudinal positioning; l) Drive; m) Slag discharge system; n) Post-combustion chamber

Waste Feeding. Continuous, efficient, and trouble-free operation of rotary kilns requires uniform metering of the waste, particularly the heating value. The mixture of solid, pastelike, and liquid wastes must be controlled so that a homogeneous, viscous melt is formed which forms a coat of slag adhering to the lining and protecting its refractory bricks against erosion and corrosion.

The charging systems of rotary furnaces must be tailored to the properties of the hazardous wastes. The feeders are located at the head end of the furnace (Fig. 106) and are designed to handle the full spectrum of solid, pastelike, and liquid hazardous wastes. Solid wastes are conveyed by a crane or an inclined bucket elevator to the charging hopper; drums are transported by the drum elevator to the charging hopper; pastelike wastes are delivered by pumps and the feed nozzle; liquid wastes arrive via lances; and fuel oil is injected into the rotary furnace via burners. The head end of the furnace can be fixed or movable and has connections for all the above feeding devices.

The most sensitive part of the feed system is the cooled feed chute, which can exhibit cracks at the welds due to thermal stress. Because the chute extrudes about 20 cm into the drum, wastes being conveyed along the rotary furnace can tumble back, wearing the bottom part of the chute.

Some wastes are delivered in drums for operational reasons. The batch size for such wastes is limited because combustion is more uniform, the more uniform the waste feed. At heating values of over 20 MJ/kg, batches should be no larger than 50 kg. If larger drums are fed into the furnace, the heat is liberated so rapidly that even the largest possible quantity of air is insufficient for complete combustion. A charging hole 60 cm in diameter and 80 cm high is sufficient for admitting 50 kg drums. The containers are either incinerated in the rotary furnace along with the waste or are discharged with the slag after the contents have been burned out. These containers are delivered by an elevator and lock; the doors have electrical interlocks so that no connection between the combustion space and the atmosphere exists at the moment when the container drops into the furnace. In newer installations, drum feed is not allowed; drums are emptied and the contents are inlet via burners.

Wastes in pastelike form are delivered to the furnace by thick-matter pumps. Liquid and viscous materials can be mixed in oblique-rotor pumps. Small solid particles cannot damage the pump because of the play between impeller and case, and occasionally they actually are reduced in size. These pumps have delivery heads of about 6 m, so that the waste is thoroughly mixed. This type of pump is proven for paint containing residual solvents.

Severely contaminated liquids can be injected through simple burners, essentially consisting of a double-walled tube with swirl chamber and nozzle; the clear diameter of such a burner is around 10 mm.

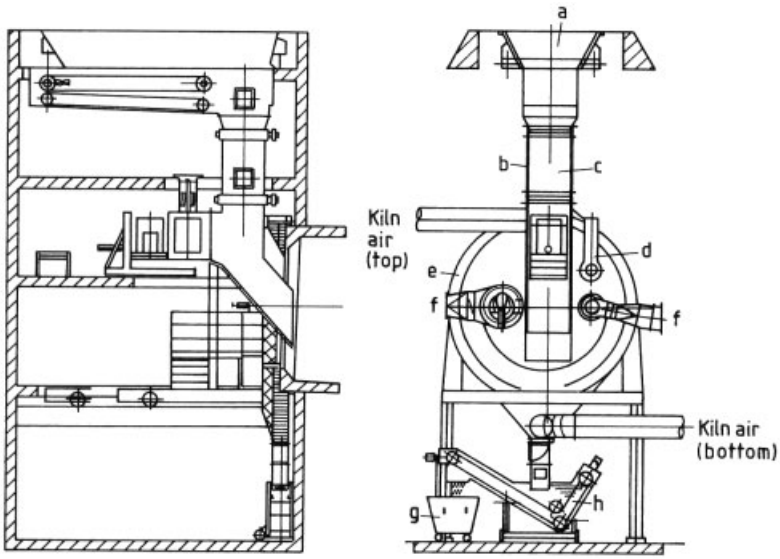

Figure 106. Charging of solid, liquid, and pastelike wastes to a rotary furnace for hazardous-waste incineration [308]
a) Solids feed hopper; b) Chute; c) Double gate (lock); d) Drum feeding; e) Lance; f) Multifuel burner; g) Container discharge; h) Wet slag discharge, furnace breeching

Some production residues, such as those from distillation, are liquid at 150 °C but solidify at room temperature. Heated or insulated pressure vessels are used for such materials; these are discharged with nitrogen from the plant system. These wastes cannot be injected through atomizing burner nozzles but flow into the furnace through a 20 mm pipe. Inside the furnace, they move along the hot wall, spread out on the surface, and readily burn. If pastelike materials contain water, droplets of water must not be allowed to impinge the hot lining, because they cause severe stress and premature wear.

Rotary Kiln. The shell of the furnace consists of welded cylindrical rings and is inclined in the direction of material flow. It has a refractory lining. As a rule, the discharge end, which experiences high thermal stress, is a double-walled cylinder with air cooling; the terminal segments are made of heat-resistant cast steel. The furnace shell is sized to be stiff in torsion. The end wall of the furnace is also lined with refractory material (Fig. 107).

Rotary kilns are 8 – 12 m long, in exceptional cases 20 m. They have diameters of 1 – 5 m. The mean combustion temperature is between 800 and 1400 °C, and residence times are ca. 60 min. Afterburners are provided to ensure complete burning of the combustion gases. Throughputs range from 0.1 to 20 t/h.

Wastes, auxiliary fuels, and combustion air are inlet through the feed end. Ash and slag exit the furnace at the other end. The furnace mounting and drive are designed to accommodate the weight of the furnace shell, wastes, and possible deposits inside the furnace. In most systems, the rotation speed can be continuously varied between 0.05 and 2 rpm.

Combustion Air Supply. The air excess is 2.0 –2.5 (to a maximum of 3.0), which is markedly higher than in grate and fluidized-bed furnaces. Rotary kilns are available in cocurrent, cocurrent twin-shell, and countercurrent designs. In a cocurrent kiln, the combustion space is bounded by the feed head, the lining, and the combustion chamber. The movement of waste and slag is in the same direction as the path of combustion air and reaction gas. In countercurrent firing, these two streams move in opposite directions and the breeching with slag discharge forms the "rear" boundary. As a rule, countercurrent operation is possible only if the combustible contents of the waste evolve no pyrolysis gases and the heating value is under ca.

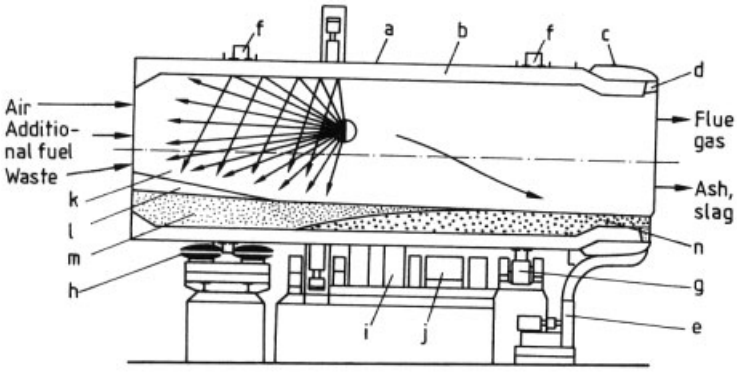

Figure 107. Stress zones in a rotary furnace [308]
a) Furnace shell; b) Refractory lining; c) Discharge ring; d) Terminal segment; e) Cooling air blower; f) Riding ring; g) Trunnion roll; h) Thrust rolls and longitudinal positioning; i) Girt gear; j) Variable-speed drive; k) Steam zone; l) Wastes; m) Combustible matter

8 MJ/kg. Chemical wastes nearly always have large amounts of carbonizable substances and higher heating values. For this reason, countercurrent operation is not common in hazardous-waste processing, though it is used for thermal cleaning of soils. Countercurrent incineration is advantageous for low-heating-value wastes with very high water contents, since the vapors from the drying zone pass directly into the combustion chamber without imposing any stress on the rest of the furnace combustion space. As a result, the wastes are burned more quickly, and a shorter furnace is adequate for complete combustion.

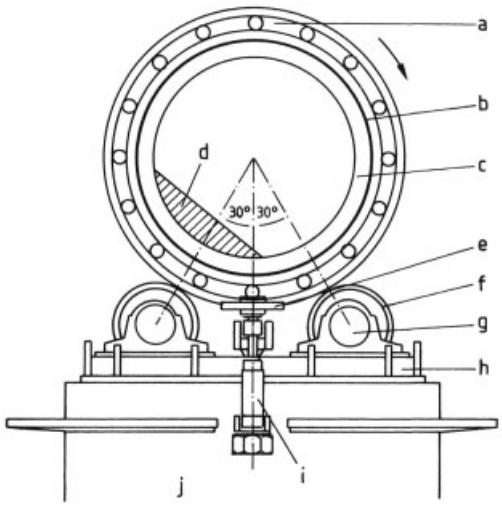

Figure 108. Cross section through rotary kiln mounting [308]
a) Riding ring (welded); b) Furnace shell; c) Lining; d) Burden; e) Thrust roll; f) Trunnion roll; g) Pillow block; h) Base plate; i) Positioning hydraulics; j) Footing

In cocurrent operation, fuel and air are inlet at the cold end, while flue gases and slag exit the furnace at the hot end. If the heating value declines, the ignition boundary moves from the feed end toward the discharge end. The ignition of the fuel can be moved back by turning on the auxiliary burners at the beginning of the shell.

Rotary furnaces are driven by a girt gear, pinion, and variable-speed motor. The rotation speed and thus the residence time of the furnace contents are controlled in this way.

Fuel feed and slag withdrawal must take place at grade level. The rotary furnace is therefore supported on a table some 4.5 m high. The riding rings (Fig. 108) bear the entire weight of the furnace including lining and charge.

If the heating value of the wastes is so low that the material does not ignite and continue to burn, supplementary fuel can be added to the feed or inlet through auxiliary burners. Liquid or pastelike wastes with high heating value can be used for this purpose, but the feed rate cannot be increased indefinitely because this would result in damage to the slag coating.

Lining. Most of the wear in rotary furnaces is due to thermal, mechanical, and chemical stress on the refractory lining. This wear results from the continual, recurrent contact of the charge with the furnace lining. No ceramic material can satisfy all requirements, so the lining must be renewed at regular intervals. Damage to the lining in the furnace feed zone is chiefly mechanical, caused by large and hard objects such as metal-

reinforced pieces of bulky waste, iron drums, and so forth [307].

Thermal stresses due to temperature cycling are produced when moisture is expelled from wastes or when pyrolysis gases are formed at temperatures up to ca. 300 °C.

Bricks with extra thermal fatigue resistance are used in this zone. The lining in the evaporation and melting zone, where the temperature is between ca. 300 and 900 °C, is attacked by solid and liquid compounds, both organic and inorganic, and also as a result of gas and vapor evolution. Hydrocarbon mixtures that ignite early can lead to abrupt and intense flame formation in the bed of material. The thermal stress is highest in the main combustion zone, where there are sporadic temperature peaks up to ca. 1500 °C; in addition, slags and dusts here react with constituents of the refractory material. Attack on the liner is also promoted by infiltration into the brick surface structure when low-melting mixtures are present. If the furnace rotation speed is low, the protective slag coating on the lining can solidify. Wastes that are difficult to ignite or oversize burn in the burnout zone of the furnace discharge end. With the radiant heat from the afterburner chamber, a high temperature profile is achieved so that the slag can be maintained in a thin liquid state. In the breech part of the rotary furnace, the refractory lining can also be attacked by abrasive metal objects in the slag. Especially dense bricks that can resist liquid slag and aggressive gaseous species are used here. The size and form of each stress zone are variable and depend on the operating parameters. The basic suitability and rating of refractories against attack by solid and liquid substances at high temperature are established by the crucible method of DIN 51069. In this test, finely divided slag is allowed to act on a hollowed-out specimen of the ceramic material. Ash melting point and application time are determined or set ahead of time. The slagging resistance is assessed by visual examination of crucible section surfaces.

A single-layer lining is generally 250 mm thick. It is made from double deep-arch bricks or standard arch bricks like those used in rotary cement kilns.

The radial furnace wall is made of ring or bonded brick masonry. Joints are kept narrow and filled with a mortar suitable to the type of brick and forming a chemical – ceramic bond. The periphery of the refractory lining must be stressed to prevent relative motion between the brick courses and the furnace shell. Expansion joints are made to allow for thermal expansion of the refractory in the longitudinal direction. The configuration, size, shape, and sealing of these expansion joints are defined case by case.

A wear-resistant 40 mm underlining of fireclay splits is a good way to enhance the emergency running properties and hold down the temperature of the steel shell. The dam at the head end and the taper in the breeching section are preferably lined with shaped bricks. A steel ring welded in the discharge zone of the drum takes the axial thrust of the masonry. Figure 109 shows a half-section of a rotary furnace with lining.

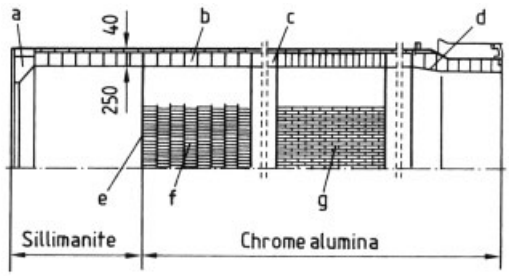

Figure 109. Partial section of rotary-furnace lining [308] a) Dam; b) Arch bricks; c) Hard fireclay splits; d) Taper; e) Expansion joint; f) Ring masonry; g) Bonded masonry

The design of the furnace lining affects its service life. In general, single-layer masonry is preferred; this is commonly sufficient even for outdoor furnaces. If thermal radiation considerations make an insulating lining indispensable, multiple courses of brick must be placed. Occasionally, two-layer masonry is laid.

Frequent startup and shutdown impairs the durability of the masonry because of the alternating stresses. Masonry repairs are necessary particularly in the transition region from the main shell to the post- combustion chamber, which is the region of maximum temperature and high gas velocity.

Post-combustion. As a rule, the residence time of the reaction gases in the furnace is insufficient for complete burning, so post-combustion chambers are employed. Secondary and sometimes tertiary air is injected into these chambers, and additional burners may also be installed there.

The post-combustion chamber has three key tasks:

1) To ensure complete combustion of the reaction gases
2) To burn liquid wastes injected directly into this chamber through burners
3) To mix the furnace and burner flue-gas streams with the secondary and tertiary air

Flue gases undergo complete combustion when the various gas streams are mixed; theymust be at roughly the same temperature. The gases must therefore have a residence time of at least 5 s in the high-temperature (950 °C) zone.

Slag deposits on the chamber lining may be a problem. If slag particles entrained by the flue gases are deposited on the vertical walls, they will build up, drop down, and plug the burner housing and slag discharge hole. This situation can be prevented with slag-repelling internals and auxiliary burners.

The gas streams can be mixed by installing ceramic mixing screens, which create turbulent flows and thus prevent bypassing. The screens can also effect preliminary removal of dust from the flue gases.

The lower part of the chamber experiences chemical attack by liquid slag droplets carried out of the rotary furnace. This zone is therefore lined with a chemically resistant material such as Alumidon 82. The upper part of the chamber can be lined with corundum, which contains 50 % Al_2O_3. The zone of the first boiler heating surfaces, where the temperature of the flue gases is declining, is lined with A_O and A_I qualities.

Because high-quality masonry is sensitive to frequent alternating thermal loads, continuous operation is recommended. For repairs in the second pass or when waste incineration is briefly interrupted, the temperature of the masonry is held at about 500 °C with fuel oil. To accomplish this, the flue-gas valve is closed and the gases are allowed to escape through an emergency stack at the end of the first pass.

The lining of the afterburner chambers with A_{Ot} bricks yields service lives exceeding 25 000 h. The shape of the post-combustion chamber can be optimized by flow tests; the same is true of the secondary and tertiary air blast points and the blast energy required.

Solid residues from combustion in the rotary kiln take the form of ash or slag, depending on the operating temperature. Ash can be discharged dry or wet; liquid slag must always be discharged via a water bath. When ash is discharged dry, the combustion residues from the furnace fall into a hopper with a water-cooled conveying trough beneath. Dry ash discharge has the advantage that no water is used and no wastewater is generated. Problems can occur in cooling the slag and in sealing the rotary furnace against the atmosphere. The dry ash can be cooled in multiple vibrating trough conveyors, but this requires the addition of water, through water nozzles located at the furnace discharge.

Liquid slag is discharged at high operating temperatures. This improves slag burnout and metal oxides contained in the hazardous wastes are better trapped in inert materials. This mode of operation places severe loads on the masonry, chiefly at the hot end. The furnace shell may be cooled with water to protect the masonry. Liquid slag drops into a quench tank where it is pelletized.

Figure 110 shows the mass and energy balance of a rotary furnace with post-combustion chamber.

9.4. Pyrolysis and Gasification

For some time, the classical incineration process has been faced with competition from several new thermal processing techniques for waste: thermoselect process, pyrolysis – combustion, fixed-bed pressure gasification, and entrained-bed gasification. Figure 111 shows the relationships of thermal processes and their end and intermediate products. Most processes end with a combustion step, in a boiler or an internal-combustion engine.

9.4.1. Principles of Waste Pyrolysis

Organic compounds decompose when heated to give volatile products and coke. The chemical reactions involved are influenced by three factors:

1) Starting materials:
   Chemical composition
   Water content
2) The reactor:

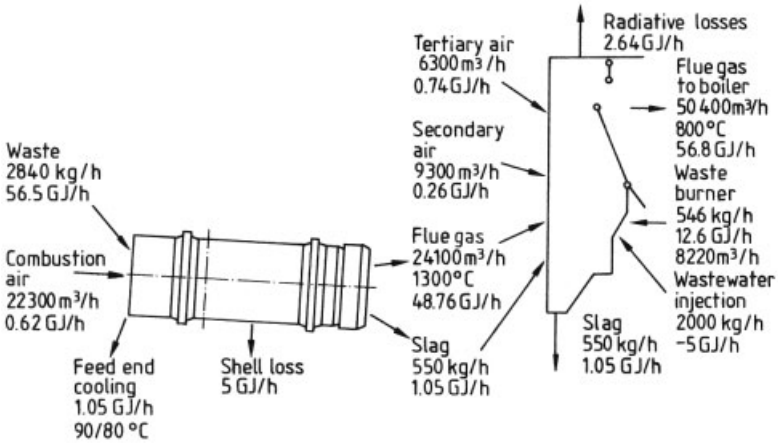

Figure 110. Mass and energy balance for a rotary furnace with post-combustion chamber used for hazardous-waste incineration
1) Waste; 2) Combustion air; 3) Feed end cooling; 4) Shell loss; 5) Tertiary air; 6) Secondary air; 7) Flue gas; 8) Slag; 9) Slag; 10) Wastewater injection; 11) Waste burner; 12) Flue gas to boiler; 13) Radiation loss

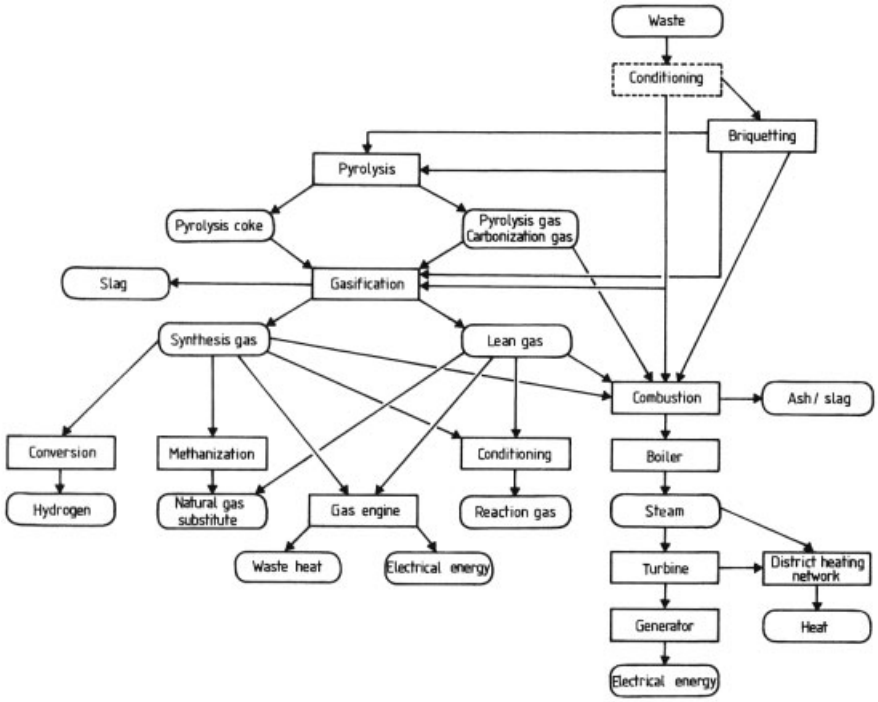

Figure 111. Options for thermal treatment of refuse

Vertical-shaft or batch reactor
Rotating tubular or fluidized-bed reactor

3) The operating conditions:
Carbonization (pyrolysis) temperature
Reaction (residence) time

As the reaction time and temperature increase, the composition of the products evolves toward more stable compounds. The pyrolytic degradation of organic materials as a function of temperature proceeds as follows:

100 – 200 °C	Thermal drying, dehydration (physical)
250 °C	Deoxidation, desulfurization; loss of water of constitution and carbon dioxide, depolymerization, onset of hydrogen sulfide evolution
340 °C	Bond breaking in aliphatic compounds, onset of evolution of methane and other aliphatic hydrocarbons
380 °C	Carbonization phase (enrichment of carbon in charge)
400 °C	Cleavage of C – O and C – N bonds
400 – 600 °C	Conversion of bituminous matter to pyrolysis oil or pitch
600 °C	Cracking of bituminous compounds to thermally stable substances (gaseous, short-chain hydrocarbons); production of aromatics (benzene derivatives) by the following hypothetical reaction scheme:
>600 °C	Dimerization of olefins (ethylene) to butylene; dehydrogenation to butadiene; diene reaction with ethylene to cyclohexene; thermal aromatization to benzene and higher-boiling aromatics

The pyrolysis process can be broken down into the following phases, which occur in succession: (1) drying phase, (2) carbonization phase, (3) gas evolution phase.

At temperatures up to 200 °C, physical drying is accompanied by the production of water vapor; this requires much energy (2270 kJ/kg water at 1 bar). If the water content is high, for example because sewage sludge, liquid manure, or liquid hazardous wastes have been added, it may be preferable to insert presses or dryers upstream of the reactor.

Carbonization takes place at 200 – 500 °C. Side groups in high molecular mass organic substances such as cellulose, proteins, fats, and plastics are cleaved. Macromolecular structures are broken down into gases, liquid organic compounds, and solid carbon. In the gas evolution phase between 500 and 1200 °C, the carbonization products are further degraded. Solid carbon and liquid organic products yield the stable gases H_2, CO, CO_2, and CH_4.

The main reactions in the gas evolution phase are the following:

$$C_nH_m \rightarrow x\,CH_4 + y\,H_2 + z\,C \quad (1)$$

$$CH_4 + H_2O \rightarrow CO + 3\,H_2 \quad (2)$$

$$C + H_2O \rightarrow CO + H_2 \quad (3)$$

$$C + CO_2 \rightarrow 2\,CO \quad (4)$$

where *m, n, x, y,* and *z* are partition coefficients.

According to Barniske, higher molecular mass hydrocarbons are decomposed to hydrogen, methane, and solid carbon [310]. Methane reacts with water [311] and is converted to carbon monoxide and hydrogen (synthesis gas). The solid carbon product reacts with water vapor or carbon dioxide to form water gas or generator gas. The pyrolysis of cellulose-containing substances, such as paper, board, wastes from kitchen and garden, and wood waste, is governed by the degradation behavior of cellulose, as illustrated in Figure 112. This process is described by the equation

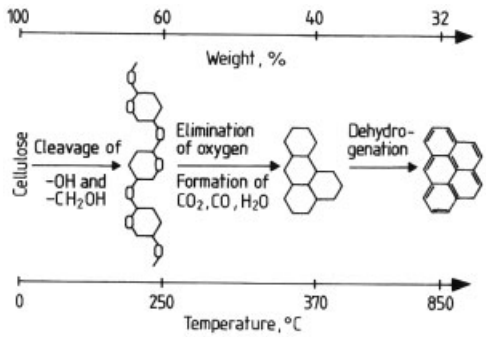

Figure 112. Cellulose degradation as a function of pyrolysis temperature

$$3\,(C_6H_{10}O_5) \rightarrow 8\,H_2O + C_6H_8O + 3\,CO_2 + CH_4 + H_2 + 8\,C \quad (5)$$

Thermal degradation of polymer materials involves the loss of side chains or substituents and yields products such as H_2O, CO_2, NH_3,

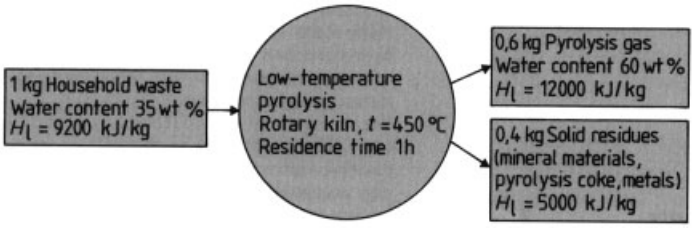

Figure 113. Material and energy balance for low-temperature pyrolysis of residential waste [313]

HCl, H_2S, CH_4, and higher aliphatic hydrocarbons. Polymers can undergo thermal depolymerization to the monomers, while polymer chains can be stochastically fragmented into fragments of various lengths. For example, paraffin wax is formed by heating of polyethylene and polypropylene. These primary decompositions are mostly accompanied by secondary reactions in which the primary degradation products can be repolymerized. At low temperatures, the reactive primary products are partly or almost wholly polymerized to tar. Some of the tar is further converted to gases and coke. At higher temperatures, the primary degradation products have a lower tendency to polymerize [312].

The pyrolysis of wastes generates a varying distribution of gaseous, liquid, and solid products.

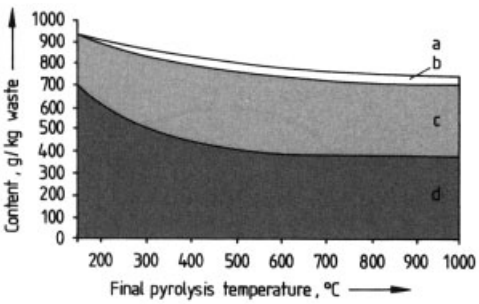

Figure 114. Composition of products from residential waste pyrolysis versus final pyrolysis temperature (see Table 27 for inlet composition) [311], [316]
a) Pyrolysis gas; b) Tar/oil condensate; c) Aqueous condensate; d) Solid residues

The end products can be:

1) Fuels or feedstocks (tars, oils, combustible gases)
2) Condensate (water) with dissolved impurities
3) Residues such as coke, metals, glass, and sand

Figure 113 shows schematically the material and energy balance for the pyrolysis of household waste. The products from pyrolysis of household waste are listed in Table 27 according to pyrolysis temperature.

The yield of liquid degradation products such as oils and tars is highest at temperatures below 550 °C. In the medium temperature range, up to 800 °C, the principal products are high-heating-value gases, with smaller quantities of tars and oils.

The yield of lower-heating-value gases is high up to about 1100 °C. When the inlet wastes have a high heating value, solid residues are melted if the temperature is increased to 1400 °C.

Figure 114 shows the product distribution from household waste pyrolysis as a function of final temperature. The quantity of solid residues declines as the final pyrolysis temperature rises to about 500 °C, then remains constant at roughly 0.4 kg/kg waste.

As pyrolysis continues, the content of volatiles in the solid residues decreases, approaching zero at around 500 °C. The carbon content decreases only slightly, to a value of ca. 18 – 20 wt % between 400 and 800 °C. Between 800 and 1000 °C, the carbon content falls to ca. 13 wt %. While the hydrogen content no longer changes appreciably above 800 °C, the CO/CO_2 ratio shifts significantly to the right in accordance with the Boudouard equilibrium, $CO_2 + C \rightleftharpoons 2\,CO$. In tests with waste between 800 and 1000 °C, where the content of organic matter was approximately constant at 34.8 – 39.8 wt % and the moisture content was

Table 27. Household waste pyrolysis products as a function of pyrolysis temperature [314], [315]

Pyrolysis temperature, °C	Pyrolysis gas, kg/kg waste	Acids, condensate, tar	Coke, kg/kg waste	Pyrolysis gas composition, vol %						Coke composition, vol%		
				H_2	CH_4	CO	CO_2	C_2H_4	C_2H_6	Volatile carbon	Bound carbon	Ash
482	0.123	0.611	0.247	5.56	12.43	30.50	44.77	0.45	3.03	21.81	70.48	7.71
648	0.186	0.186	0.592	16.58	15.91	30.49	31.78	2.18	3.06	15.05	70.67	14.28
815	0.237	0.597	0.172	28.55	13.73	34.12	20.59	2.24	0.77	8.13	79.05	12.82
926	0.244	0.587	0.177	32.48	10.45	35.25	18.31	2.43	1.07	8.30	77.23	14.47

between 17.8 and 21.2 wt %, the following distributions were obtained:

800 °C: 23.0 vol % CO_2, 13.0 vol %CO
900 °C: 18.3 vol % CO_2, 20.5 vol % CO
1000 °C: 12.0 vol % CO_2, 29.0 vol % CO

For a roughly constant waste composition, the methane concentration in the pyrolysis gas decreases with increasing process temperature. The CH_4 content decreases from 9.5 vol % at 800 °C to 3.5 vol % at 1000 °C. The formation of oil and tar condensate reaches a maximum between 350 and 500 °C. At higher temperatures, cracking reactions occur and reduce the quantity of tar.

The amount of aqueous condensate depends on the water content of the waste and, to a lesser extent, on the water liberated by degradation during pyrolysis. The quantity of condensate increases up to ca. 300 °C because of the moisture in the waste (32.1 wt %), then remains virtually constant at higher temperatures. It drops if water-gas reactions begin.

The pyrolysis gas mainly consists of simple, low molecular mass compounds. The combustible species are hydrogen, carbon monoxide, methane, and aromatic and higher hydrocarbons. Noncombustible carbon dioxide is produced by the decomposition of oxygen-containing compounds and the combustion of organic compounds. The quantity of pyrolysis gas increases with the pyrolysis temperature, from 70 g/kg household waste at 150 °C to ca. 230 g/kg at 1000 °C. The composition of the gas is shown as a function of process temperature in Figure 115.

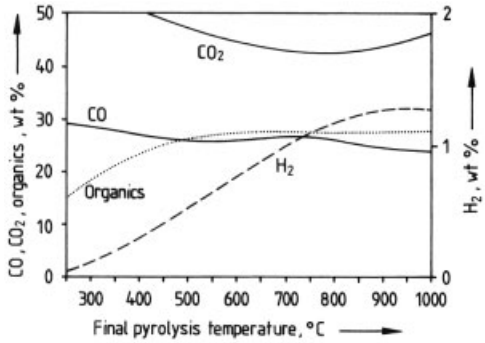

Figure 115. Composition of pyrolysis gas versus final pyrolysis temperature [311]

The material balance and gas composition plotted in Figures 114 and 115 are for household waste with the following characteristics [311]:

Lower heating value (inferior calorific value) H_1, MJ/kg	7.2
Water content, wt %	32.1
Ash, wt %	34.3
Carbon, wt %	19.0
Hydrogen, wt %	6.2
Oxygen, wt %	38.7
Nitrogen, wt %	1.1
Sulfur, wt %	0.1
Chlorine, wt %	0.6

The final pyrolysis temperature is thus the governing parameter for the heating value of the pyrolysis gas. While H_2 and CO with their higher heating values (superior calorific values) of $H_h = 12.7\,MJ/m^3$ and $12.6\,MJ/m^3$ (here and in what follows, gas volumes are at STP) are of the same order, the values for CH_4 and C_2H_4 are factors of roughly 3 and 5 higher. If a high-energy-content gas is desired, the pyrolysis reactions must be carried so that large amounts of

methane and higher hydrocarbons are produced [317].

The effect of the process temperature on the gas composition has been confirmed for the pyrolysis of paper (Fig. 116) [318]. While the concentration of carbon monoxide in the pyrolysis gas remains approximately constant with rising temperature, the content of carbon dioxide declines. CH_4 and C_nH_m have their maximum at ca. 550 °C, but decrease above this temperature due to elimination of hydrogen.

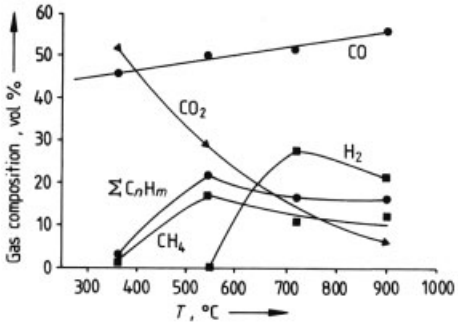

Figure 116. Influence of process temperature on gas composition in the pyrolysis of paper [318]

Figure 117 shows how the final pyrolysis temperature of paper affects the heating value of the pyrolysis gas. The highest heating value is achieved at 700 °C. At higher temperatures, the heating value falls off as the CH_4 and C_nH_m concentrations decrease.

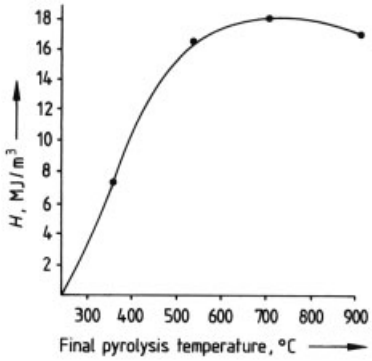

Figure 117. Heating value of pyrolysis gas as a function of final temperature in the pyrolysis of paper [318]

9.4.2. Principles of Gasification of Waste

In gasification, an agent is added to convert carbonaceous waste constituents to gaseous fuel and ash or slag. After pyrolysis, a reactive gas is admitted to convert the coked residues to further gaseous products. This is thus a continuation of the pyrolysis process. The residual carbon content of the pyrolysis coke in the red-hot bed is oxidized at temperatures above 800 °C with a substoichiometric quantity of oxygen.

Steam, carbon dioxide, oxygen, or air can be used as the gasifying agent. Like pyrolysis, gasification is both a process in itself and a subprocess of combustion, or of combined pyrolysis and combustion processes such as the Thermoselect process. The products generated by gasification depend on the gasifying agent used (lean gas or synthesis gas). The reaction energy for gasification is produced by partial combustion of the organic material in the gasification reactor. The process is described by the following equations, depending on which of the gasifying agents (oxygen, air, steam, hydrogen, carbon dioxide) is employed:

Oxidation (exothermic):

$$C + O_2 \rightarrow CO_2 \tag{6}$$

$$C_xH_y + (x + {}^1/^4\, y)\, O_2 \rightarrow x\, CO_2 + {}^1/^2\, y\, H_2O \tag{7}$$

$$C + 2\, H_2 \rightarrow CH_4 \tag{8}$$

Boudouard reaction (endothermic):

$$C + CO_2 \rightleftharpoons 2\, CO \tag{9}$$

Water-gas reaction (endothermic):

$$C + 2\, H_2O \rightleftharpoons CO_2 + 2\, H_2 \tag{10}$$

$$C + H_2O \rightleftharpoons CO + H_2 \tag{11}$$

$$C_xH_y + x\, H_2O \rightarrow x\, CO + (x + {}^1/^2\, y)\, H_2 \tag{12}$$

The fuel bed is brought to red heat by partial combustion of the carbon (Eqs. 6 and 7) and reacts with the carbon dioxide formed (Eq. 9). The higher the temperature in the fuel bed, the more carbon monoxide is formed. The CO/CO_2 ratio is represented by the Boudouard equilibrium in Figure 118. The water-gas reaction behaves in a similar way: when steam is mixed with the gasifying air, CO_2 and hydrogen are formed at low

temperatures (Eq. 10) while CO and hydrogen are formed at high temperatures (Eq. 11).

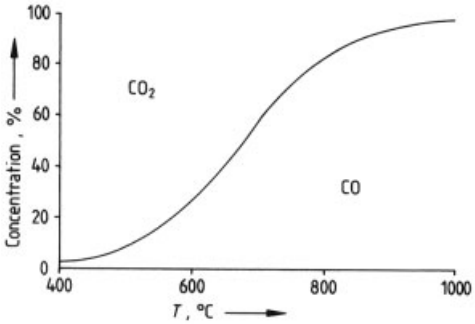

Figure 118. Equilibrium of the Boudouard reaction at atmospheric pressure [319]

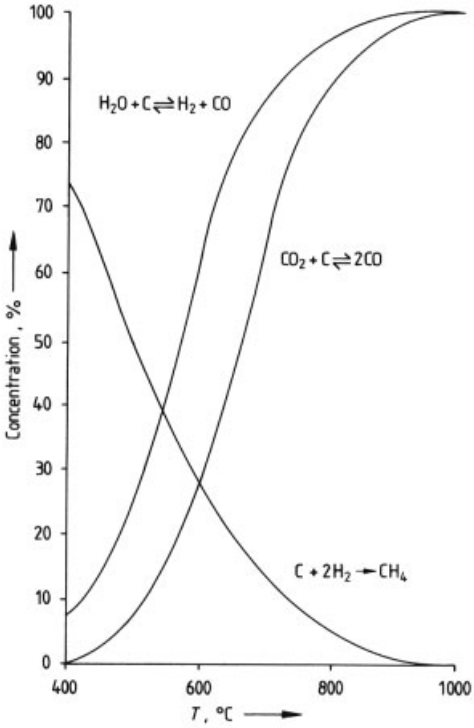

Figure 119. Equilibrium of the carbon gasification reaction at 1 bar as a function of temperature [319]

Gasification with hydrogen (Eq. 8) is also exothermic, but this method is not practiced on an industrial scale for reasons of cost. The partial equilibrium reactions (9) and (11) are plotted as functions of temperature in Figure 119 and pressure in Figure 120. The equilibrium gas composition when carbon is gasified with steam is shown in Figure 119. The gasification or reaction rate depends not only on the temperature but also on the porosity, pore diameter, and internal diffusion of the fuel bed. This is why pyrolyzed (i.e., degassed) waste gasifies better than unconditioned waste. The same applies to lumpy material in comparison with loosely piled material. Lumpy (briquetted) material allows a better flow of gas through the fixed bed in the vertical-shaft reactor; at the same time, the formation of gas preference paths–which would lead to incomplete gasification–is prevented. When pyrolysis and gasification take place in a single reactor, the heat required to sustain the dual process is supplied by the partial combustion of carbon in the lower section of the gasification reactor.

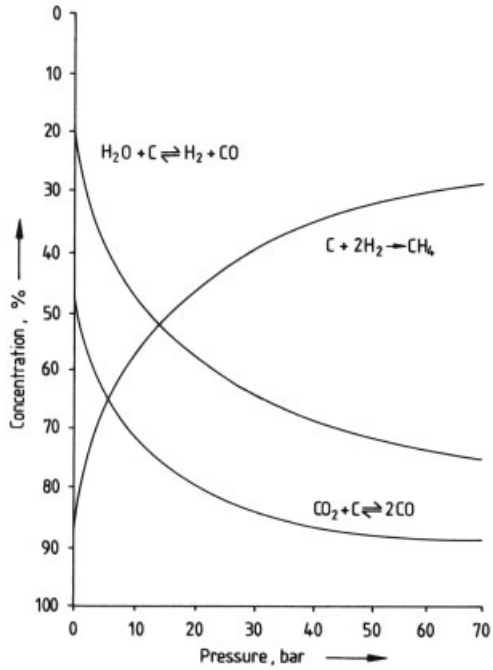

Figure 120. Equilibrium of the carbon gasification reaction at 700 °C as a function of pressure [319]

The gaseous products formed in gasification depend on the gasifying agent and the pyrolysis temperature. The gasification products can be classified by heating value:

Lean gas :	$H_1 < 12\,500\,\mathrm{kJ/m^3}$
Rich gas :	$H_1 > 12\,500\,\mathrm{kJ/m^3}$

or according to composition and gasifying agent:
Generator gas (gasifying agent: air)

CO	21 – 30 %
H_2	11 – 18 %
CO_2	5 – 11 %
N_2	47 – 56 %
H_1	5040 kJ/m^3

Water gas (gasifying agent: steam)

CO	40 %
H_2	50 %
CO_2	5 %
N_2	4 %
H_2O	1 %
H_1	12 600 kJ/m^3

Figure 121 shows the equilibrium gas composition for the gasification of carbon with steam.

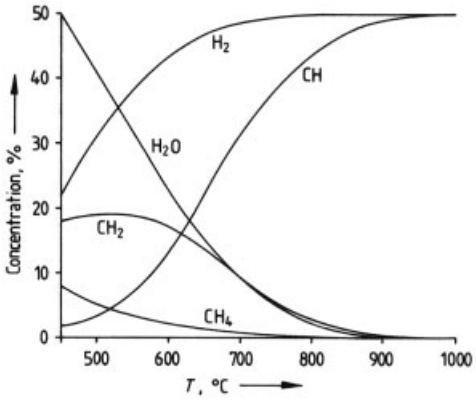

Figure 121. Equilibrium gas composition in gasification of carbon with steam [319]

A steam – air mixture can also be used for gasification, yielding a "mixed" gas whose composition and heating value fall between those of generator gas and water gas and depend on the mixing ratio of the gasifying agent.

Gasification processes operating at high temperatures (>1600 °C) give a higher gas yield, albeit with a lower heating value, than processes running at moderate temperatures. The heating value of the gases from waste pyrolysis is higher than that from gasification with air:

Pyrolysis gas:	12 000 – 16 000 kJ/m^3
Gas from air gasification:	ca. 5000 kJ/m^3

The latter is of roughly the same order as generator gas from coke gasification (H_1 = 5195 kJ/m^3). If the waste is gasified with oxygen as in the Thermoselect process, the heating value rises to about H_1 = 10 000 kJ/m^3.

Although generator gas has a lower H_1 than pyrolysis gas, the volume of gas per unit mass of waste is larger. The noncombustible substances are converted to ash or slag, depending on the process temperature. As a rule, the melt obtained at high gasification temperatures is pelletized in a wet deslagger. This largely sterile pelletized slag product can be used as aggregate in the construction industry. The solid residues from low-temperature gasification processes are similar to those from combustion, with high ash contents and low carbon contents.

9.4.3. Processes

9.4.3.1. Pyrolysis – Combustion Process

The pyrolysis – combustion process, developed by Siemens KWU for the treatment of household waste [320], combines pyrolysis and high-temperature combustion. The objective is to convert waste into recoverable substances with virtually no residue.

The following process steps are combined: pyrolysis of the feed; recovery of the valuable substances iron, nonferrous metals, and glass/stones from the solid raw material inlet to pyrolysis; and subsequent high-temperature combustion of all the remaining fine carbonaceous residue together with the pyrolysis gas. The organic constituents are degraded, while the noncombustible inorganic material in the residual fines are melted. The melt solidifies in granular form in a water bath; this product is intended for the same uses as the melting-chamber pellets from coal-fired power plants. In the pyrolysis-combustion process, the only residues destined for disposal are those from the flue-gas cleaning system; all other products can be recovered.

Description of the Process. As shown in Figure 122, the pyrolysis – combustion process can be broken down into the following steps:

1) Waste treatment
2) Pyrolysis
3) Treatment of residues

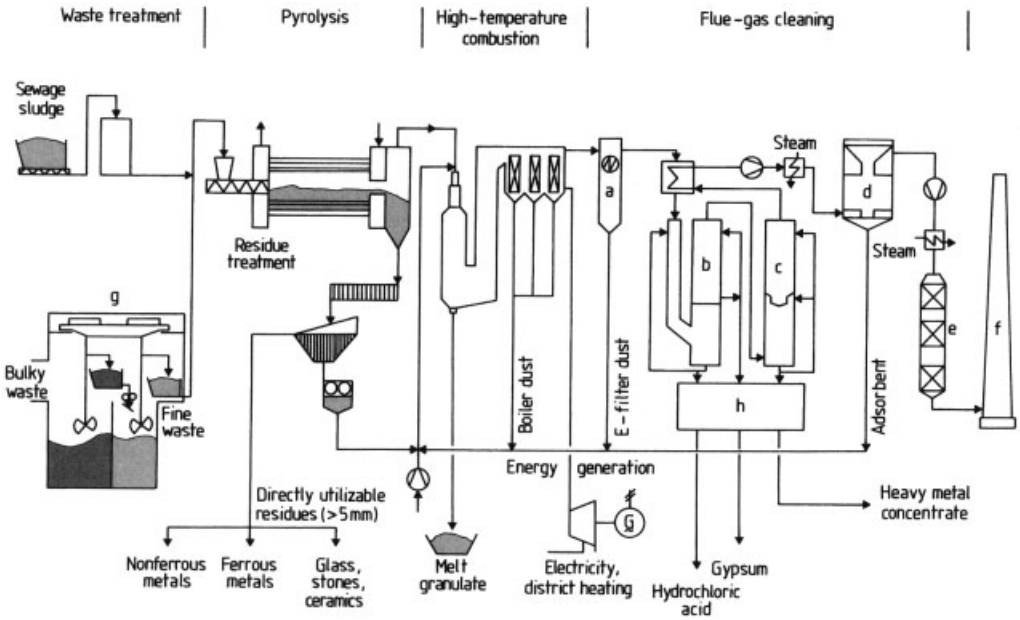

Figure 122. Flowsheet of a pyrolysis – combustion plant [320]
a) Electrostatic filter; b) HCl scrubber; c) SO_2 scrubber; d) HOK fixed-bed reactor; e) $DeNO_x$ reactor; f) Stack; g) Refuse bunker; h) Wastewater treatment

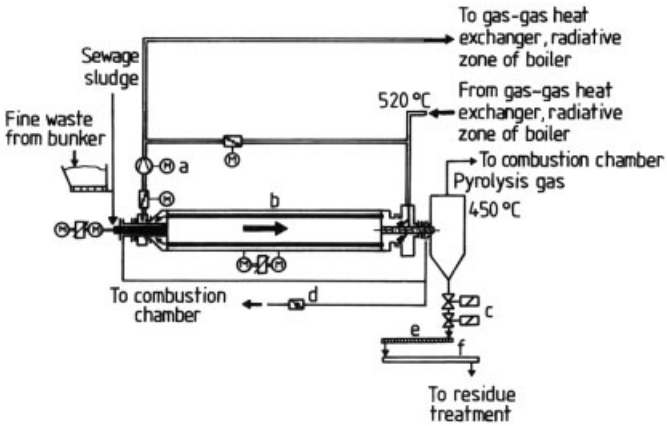

Figure 123. Conversion drum used in the pyrolysis – combustion process
a) Hot gas fan; b) Conversion drum; c) Discharge lock for residues; d) Drum seal suction purge; e) Vibrating cooling trough; f) Collecting trough

4) High-temperature combustion
5) Flue-gas cleaning
6) Energy production

In the treatment step, the waste stored in the refuse bunker is cut by rotary shears, to a maximum size of ca. 200 mm, and then conveyed into the conversion drums by a pusher screw. Sewage sludge can be added in the screw conveyor. The waste mixture is pyrolyzed in the conversion drum; the final temperature is 450 °C. Two homogeneous fuels are produced from the organic content of the waste: conversion gas and fine carbonaceous residue. Both fuels are burned at about 1300 °C in the high-temperature combustion chamber. The conversion drum (Fig. 123) is a rotating kiln supported at both ends. For a throughput of 5 t/h, the drum is 22 m long and

2.8 m in diameter. It is closed at either end with tube sheets into which straight heating pipes are welded.

In drum rotating at 2 rpm and inclined at 1.5°, the average residence time of the waste is ca. 1 h. Flue gas from the boiler downstream of the high-temperature chamber flows through the heating pipes in countercurrent to the waste; the inlet temperature of the gas is 520 °C and the outlet temperature at the end of the drum is 250 °C. The heat from this gas is essentially transferred to the waste by conduction between the waste constituents and the hot surfaces of the heating pipes. If the heating surface is ca. 120 m^2/t, ca. 210 kWh/t is required to heat waste with a moisture content of ca. 25 % to the final temperature of 450 °C.

To maintain reducing conditions in the conversion drum, air leakage is avoided when the waste is charged into it. The pusher screw conveys waste into the pyrolysis drum continuously. The degressive screw pitch provides a small amount of compaction and forms the waste into a plug. This waste plug seals the reaction space of the conversion drum against the surroundings.

The pyrolysis drum is operated at a slight vacuum, while the conversion drum is unpressurized on the hot-gas side. The pressure in the reaction space at the cold end of the drum is nearly atmospheric; a slight vacuum of 2 mbar prevails at the hot end.

All the parts of the conversion drum, from the pusher screw at the cold end of the drum to the rotating shell, the internal heating pipes, and the discharges for conversion gas and residue, are made of steel. Under service conditions, the materials subject to wear exhibit a corrosion rate of less than 0.1 mm/a. Conversion residue and pyrolysis gas are produced from the waste mixture in the conversion drum. A carbonaceous dust is obtained by screening the pyrolysis residue in the residue treatment step. Pyrolysis gas and this dust residue are burned at ca. 1300 °C in the high-temperature combustion chamber (Fig. 124). The temperature in this chamber is controlled by preheating of combustion air and off-gas recirculation. The high-temperature combustion chamber consists of a vertical cylindrical combustion space lined with refractory and having no internals; it is equipped with a multi-fuel burner.

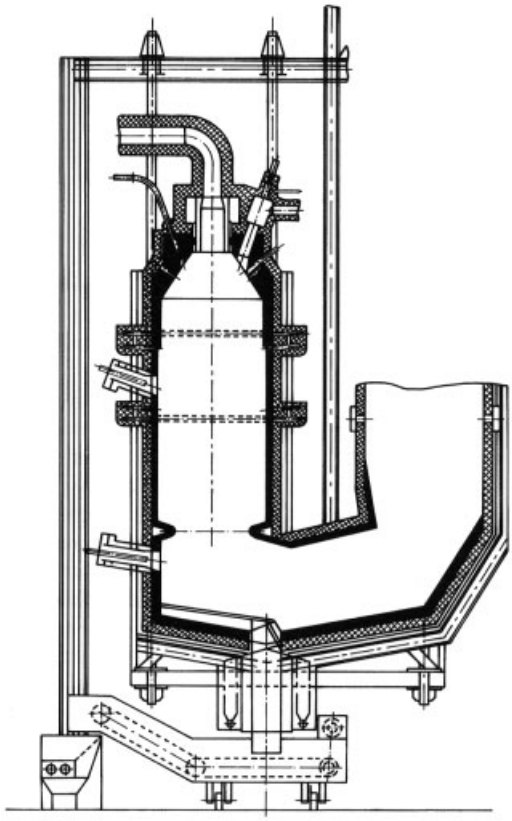

Figure 124. High-temperature combustion chamber used in the pyrolysis – combustion process [320]

The conversion gas is led from the conversion drum directly to the burner. The carbon-containing dust passes through an interim storage stage after residue treatment and is metered into the combustion chamber according to operational requirements. With multistage air feed and the resulting stagewise combustion, NO_x production is minimized. In the first stage, the off-gas temperature is established at ca. 1000 °C by controlled admission of air. The stepwise admission of secondary and tertiary air raises the temperature to ca. 1300 °C in the combustion chamber. The residence time of the off-gas in the temperature range 1000 – 1300 °C is >2 s.

The fine residue has a high content of inert materials, such as sand, glass dust, and abraded porcelain particles. The melting point of this material lies between 1100 and 1150 °C.

At the operating temperature of ca. 1300 °C, a fluid slag resembling glass or basalt is obtained. To protect the refractory lining, the outside walls

of the combustion chamber are slightly cooled. A protective layer of solidified slag is produced at the lining surface temperature of ca. 1050 °C, and the liquid slag drains off over this layer. The service life of the combustion chamber lining is extended in this way. At the bottom of the combustion chamber, the slag flows through a discharge dip tube into the water-filled wet deslagger. The residual carbon content of the slag is <0.1 %.

The hot (1250 – 1300 °C) off-gas exits the high-temperature combustion chamber and flows into the downstream waste-heat boiler. Virtually complete combustion is achieved with the combustion of the largely homogeneous fuels, conversion gas and carbonaceous fine residue, in the high-temperature combustion chamber. Off-gas cleaning resembles the wet scrubbing process downstream of conventional waste incinerators. The principal components of the off-gas treatment system are an electrostatic filter for dust collection, a spray dryer to concentrate the scrubber slurry, a fabric filter to collect fine dust, a two-stage scrubber system, a $DeNO_x$ catalytic unit, and a final cleaning stage.

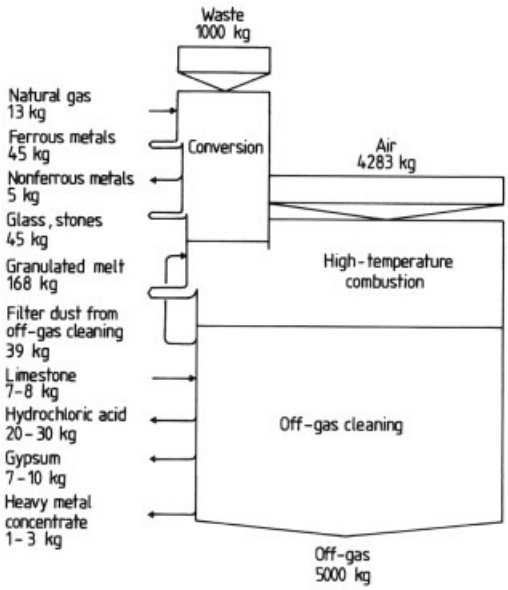

Figure 125. Mass balance of the pyrolysis – combustion process

An induced-draft fan downstream of the scrubber system generates a slight vacuum throughout the system, conveying the off-gases through the cleaning system. A fixed-bed filter or entrained-bed reactor is used for final cleaning. In the former case, the off-gas flows through a packed bed of open-hearth coke, which adsorbs organic contaminants and heavy metals.

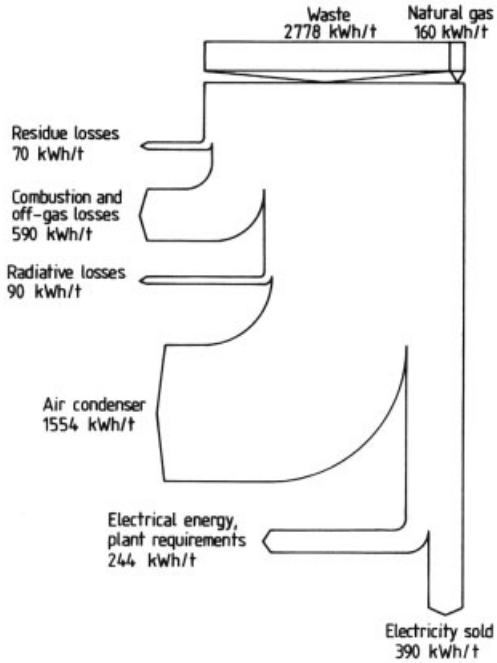

Figure 126. Energy balance of the pyrolysis – combustion process

If an entrained-bed reactor is employed, a mixture of lime and activated carbon is injected into the off-gas stream and then collected in the downstream fabric filter. As the gas flows through this filter layer, residues of organic contaminants and heavy metals are removed. Dioxin and furan emissions after this stage, as well as heavy-metals emissions, are below the limits set forth in the 17th Federal Pollution Control Regulation (17. BImSchV). Ways to separate off-gas cleaning residues into recoverable products are under study.

Hydrochloric acid and gypsum can be produced in a modified off-gas cleaning system with a distillation tower and gypsum dewatering, as in conventional incineration processes.

The steam from the waste-heat boiler can be utilized to generate electricity or in a cogeneration scheme. After plant requirements are met, the surplus electrical and/or thermal energy can

be fed into the distribution grid. Figures 125 and 126 show the mass and energy balances.

Emissions and Products. The fine dust produced in pyrolysis and the conversion gas are burned with a slight air excess. Measurements in a pilot plant have shown operating values to be under the limits of the 17th Federal Pollution Control Regulation by a wide margin.

The pyrolysis – combustion process discharges no wastewater; wastewaters are concentrated in the spray dryer or, if hydrochloric acid and gypsum recovery are included downstream, in an evaporator. The steam is led to the stack along with the stack gases.

The solid products are:

1) Ferrous and nonferrous metals as well as glass and ceramics from residue treatment
2) Pelletized melt from the combustion chamber (see also Table 30)
3) Off-gas cleaning residues, possibly including hydrochloric acid and gypsum

The residue for landfill disposal is 0.1 – 0.3 % of the inlet waste as off-gas cleaning residues, which must be treated as hazardous waste. If hydrochloric acid and gypsum are not recovered, the quantity rises to 2.4 %.

The iron, nonferrous metals, and glass/stone fractions are virtually free of carbonaceous dust. Paints and plastic coatings on metals are removed from the substrate by virtue of the process conditions in the conversion drum, so that ferrous and nonferrous metal scraps, the latter containing over 70 % aluminum, are obtained.

The glass/ceramic fraction can be utilized in civil engineering, highway construction, or landscaping. The pelletized product from the combustion of conversion gas and carbonaceous fines can be used without further treatment.

The pelletized melt and the glass/ceramic fraction from coke processing can be used in bound or unbound subbases for highway construction; in landscaping, for example when noise-protection berms are built or when fill is needed in land reclamation projects; in the construction industry; and as aggregate.

9.4.3.2. Thermoselect Process

Description of the Process. The standard Thermoselect plant consists of two parallel lines each with a capacity of 10 t/h, a common hopper, and a shared process water-treatment system. Modular design permits quick erection of the plant and replacement of components, for example the bottom section of the high-temperature reactor–the component that experiences the most severe thermal stresses.

The process breaks down into the following stages (Fig. 127):

1) Compaction of waste to 10 % of its original volume
2) Pyrolysis in a tunnel-type reactor at temperatures of up to 600 °C
3) Gasification with oxygen in the high-temperature gasifier
4) Discharge of the melt via the homogenization reactor
5) Shock cooling of the synthesis gas to under 70 °C in the quench, with separation of water, dust, and heavy metals
6) Acid wet scrubbing to remove HCl, HF, and heavy metals
7) Sulferox scrubbing to remove sulfur compounds
8) Basic scrub and cooling of gas to 5 °C to remove water
9) Final cleaning of synthesis gas in the gas engine
10) Cleaning of process water, including evaporation

Waste Compaction. The waste, delivered with a low density, is compacted to about 10 % of its original volume in a press. As a result, moisture is evenly distributed in the waste and the residual air content in the compacted waste is reduced. Compaction of the waste promotes heat transport and is a prerequisite for the small dimensions of the waste-treatment plant.

The compactor has a gas-tight connection to the pyrolysis tunnel, afforded by the heat shield. It is mounted on rollers and prestressed with a hydraulic cylinder in order to offset strains due to heating of the pyrolysis tunnel.

The interior of the press–floor, walls, and roof–is lined with 20 mm wear plates of hard metal alloy.

Heat Shield. The heat shield is located between the compactor discharge and the pyrolysis tunnel inlet. It protects the compactor against the temperatures in the pyrolysis tunnel.

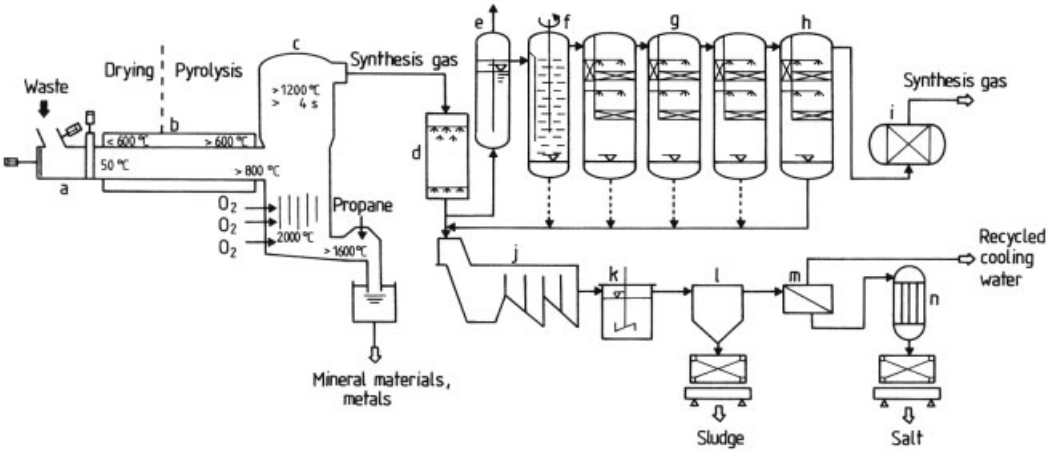

Figure 127. Flowsheet of the Thermoselect process [320]
a) Compactor; b) Pyrolysis tunnel; c) High-temperature reactor; d) Shock cooling; e) Surge tank; f) Acid scrub; g) Sulferox scrubbing; h) Gas cooling; i) Activated-carbon filter; j) Solids separation; k) Neutralization, flocculation, precipitation; l) Sedimentation, filtration; m) Reverse osmosis; n) Evaporation

Pyrolysis Tunnel. The pyrolysis tunnel (Fig. 128) located between the heat shield and the high-temperature reactor, is a double-walled reactor, rectangular in cross section. At intervals of roughly 5 – 10 min, waste is pushed through the interior of the tunnel in steps. Heat is supplied from outside through the inner wall of the tunnel, the space exterior to this wall being heated in countercurrent to the waste. The heating gas is heated in a combustion chamber by burning synthesis gas generated in the process.

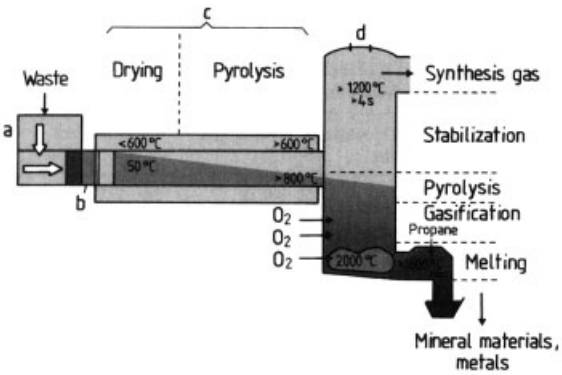

Figure 128. Thermal section of a Thermoselect plant [319]
a) Compactor; b) Heat shield; c) Pyrolysis tunnel; d) High-temperature reactor

The compacted masses of waste in the pressure-tight pyrolysis tunnel (similar in design to a pusher furnace) form a gas-tight plug. In the externally heated pyrolysis tunnel, the wastes are dried and then pyrolyzed as they pass along. In addition, heat is supplied by radiation from the downstream gasification reactor.

The pyrolysis tunnel is tapered in the process direction and has gas-discharge ducts on the inside. In this way, material transport and the removal of pyrolysis gas in the process direction is promoted. The pyrolysis tunnel is sealed gas-tight to the compactor at one end and is open to the high-temperature reactor at the other. It is mounted on rollers to compensate for thermal stresses. The lumpy pyrolysis coke consists of partly pyrolyzed organic material together with inorganic mineral and metallic residues.

High-Temperature Reactor. The partly pyrolyzed waste at ca. 600 °C, is discharged from the pyrolysis tunnel into the gasifier (Fig. 128) with the introduction of new packets of waste. The pyrolysis gas from the pyrolyzer also flows into the high-temperature gasifier.

The pressure generated by outgassing causes the briquetts to disintegrate, and the fragments are spread through the reaction space. In this way, an irregular, gas-permeable bed is created. The depth of the bed is held constant by process control of the gasification reactor.

The gasifier is a vertical-shaft reactor with an inlet opening for the partly pyrolyzed material and a gas discharge in the upper section for removal of synthesis gas. The reaction space is

bounded below by a base through which the melt is discharged to the homogenization reactor and thence to the water bath for pelletizing. The reactor has refractory lining, insulation, and cooling as appropriate to the operating temperatures.

In the lower section it is fitted with a ring of burners supplied with synthesis gas or natural gas and, in normal operation, with preheated oxygen. The oxygen serves as gasifying agent, reacting exothermically with the carbon content still present in the lower part of the bed to form carbon dioxide. The ascending CO_2 is reduced to CO as it passes through the carbon-containing bed. Ideally, this process goes to the Boudouard equilibrium.

The water-gas reaction takes place in the presence of an excess of hot water vapor, yielding carbon monoxide and hydrogen. Organic compounds are cracked.

Due to the exothermic gasification of carbon, the temperature in the lower section of the gasifier is ca. 2000 °C. All metallic and mineral constituents of the waste are therefore melted. The gases ascending in countercurrent in the reactor bring about residual pyrolysis in the top part of the bed. The most important parameters for the processes occurring in the gasification reactor are the melt flow, the height of the column of material in the reactor, and the outlet temperature and composition of the synthesis gas.

The temperature needed to maintain a fluid melt is generated by the combustion of carbon with oxygen. If this heat is not sufficient, gas can be supplied through the burners. If the synthesis gas temperature falls below 1200 °C at the reactor outlet, the oxygen quantity can be adjusted. Part of the inorganic matter vaporized by the high temperatures in the bottom of the reactor and entrained in the synthesis gas con-denses on the reactor walls and flows downwards.

Above the outlet of the pyrolysis tunnel is the stabilization zone, where gases from pyrolysis and gasification are mixed and heated to at least 1200 °C. Chlorinated and long-chain hydrocarbons are destroyed. After a residence time of ca. 4 s, the gas exits the high-temperature reactor and passes on to the shock cooling (quench) stage.

Homogenization Reactor. The melt from the high-temperature reactor flows into the melt bath of the homogenization reactor. This melt can contain carbon and impurities, which are gasified if the partial pressures of oxygen and water vapor are sufficiently high. The combined-fuel burners are arranged above the melt, pointing opposite the flow direction, so as to induce thermal turbulence in the melt. By virtue of its higher density, the metallic melt flows along the reactor bottom under the mineral melt. The melt flows over the discharge sill into a refractory-lined trough leading to the granulation system.

All gaseous reaction products flow into the high-temperature reactor.

Synthesis Gas Cleaning. Pyrolysis in the tunnel reactor and gasification with oxygen in the gasifier generate ca. 1000 m^3 of synthesis gas per tonne of waste. The synthesis gas exits the reactor at ca. 1200 °C. Its main components are carbon monoxide and hydrogen, which can be used for energy production, as well as water vapor and carbon dioxide. The gas is comparable with synthesis gas from other pyrolysis and gasification processes (see Table 28).

Sulfur inlet with the waste is present in the synthesis gas mainly as H_2S. Ammonia is unstable at 1200 °C, decomposing to nitrogen and hydrogen. Metals and metal compounds may also be present in gaseous form, depending on the vapor pressure.

The gas cleaning system has the task of removing these contaminants and collecting solid particles (ash, soot, and graphite). This system consists of the following units (Fig. 127):

1) Gas quench
2) Surge tank
3) Acid scrubbing
4) Sulferox scrubbing
5) Gas cooling
6) Adsorption

1) Gas quench: The crude gas from the gasifier flows into the first stage of the two-stage gas quench, where it is cooled from ca. 1200 °C to <90 °C against water. The purpose of this quench is to prevent the de novo synthesis of dioxins and furans that takes place when gases containing chlorine and carbon are cooled slowly.
   The crude gas flows through the quench radially; well-defined quantities of water are delivered from a radial ring slot. The formation of deposits of solid particles is prevented by a water-cooled ram.

Table 28. Composition of Thermoselect synthesis gas compared with other gasification processes

	Generator gas	Water gas	Thermoselect
CO	21 – 30 %	40 %	32 – 47 %
H_2	11 – 18 %	50 %	29 – 37 %
CO_2	5 – 11 %	5 %	19 – 29 %
N_2	47 – 56 %	4 %	ca. 2 %
H_2O		1 %	ca. 1 %
H_1	5040 kJ/m^3	12 600 kJ/m^3	up to 10 000 kJ/m^3
Gasifying agent	air	steam	oxygen

2) Surge tank: The crude gas after removal of solids and acidic constituents flows into the surge tank, which safeguards the downstream unit against overpressure. This protection consists of a submerged branching, which directs gas to the flare if the pressure becomes excessive.
3) Acid scrubbing: The cooled crude gas enters the rotary scrubber, where water is sprayed in to eliminate dust and soot particles as well as salt and acidic species such as HCl and HF. The water discharge from the quench and rotary scrubber dips into the sedimentation vessel. Hydrochloric acid is metered into the scrubber from a supply tank to hold the pH constant.
   In the second scrubber, of packed-tower type, additives are introduced to eliminate the residual fine solids with the acidic contaminants. The process water is led through a filter to the water-treatment stage. The filter cake is recycled to the waste feed end together with the solids from the sedimentation vessel.
   The sedimentation vessel consists of three chambers separated by overflow weirs. In the first, coarse solids (also containing carbon) settle out. A bucket elevator conveys the solids from the chamber into circulating containers, which carry the solids to the waste inlet end. In the second and third chambers – the second being provided with lamellae to promote sedimentation – finer particles settle out and are pumped to the water-treatment system for filtration.
   This sedimentation vessel also serves as a water supply tank for the quench water loop and acid scrub.
4) Sulferox scrubbing: Hydrogen sulfide in the gas is converted to sulfur and hydrogen in the Sulferox scrubber. For this purpose, the gas is led through a water-irrigated packed tower in cocurrent with iron chelate.
   The trivalent iron in the chelate is reduced to divalent iron by H_2S with the formation of elemental sulfur and hydrogen. A substream is withdrawn from the water loop and pumped into the regeneration stage, whereby the heat of reaction is removed via heat exchangers.
   Divalent iron in the sulfur-containing iron chelate solution is oxidized to trivalent iron by admitting air to the regeneration/sedimentation tank. The sulfur sinks downward and forms a sulfur-chelate mixture, which is pumped to a vacuum filter that continuously delivers a sulfur cake.
   The filtrate is split into two substreams, of which one is pumped back to the regeneration tank while the other goes to the collecting tank of the process-water treatment system.
   The recycle from the regeneration tank is fed back to the scrub-liquor loop of the Sulferox scrubber. Water mist is eliminated from the gas by a packing; the water goes to the cooling stage.
5) Cooling: Any traces of heavy metals and water are eliminated by two-stage low-temperature drying. The gas stream is cooled to 36 °C against cooling water in the first stage, to 5 °C against chilled water in the second. The elimination of water increases the heating value of the synthesis gas.
   After mist elimination, the gas is next reheated against warm water in a heat exchanger.
6) Adsorption: The remaining gaseous contaminants are adsorbed in an activated coke filter. The spent coke is recompacted with waste and pyrolyzed. The purified synthesis gas can be employed for heating the pyrolysis

tunnel or for energy production in an internal combustion engine or gas turbine.

The mass and energy balances are shown in Figures 129 and 130.

Process-Water Treatment. A water treatment unit is provided for treating all waters from gas cleaning (Fig. 131). Two main wastewater streams must be distinguished: the virtually solids-free main stream from gas cooling and drying, and the solids-laden stream from the sedimentation tank, into which wastewater from the quench and acid scrubbing stages have been collected.

The solids (graphite and soot particles, heavy metals, and salts) are sedimented in the sedimentation tank. The purified water is pumped back to the gas cleaning unit.

All process water is subjected to treatment in the following stages:

1) Neutralization and precipitation
2) Flocculation and sedimentation
3) Microfiltration
4) Reverse osmosis
5) Evaporation

1) Neutralization and precipitation: In this stage, the pH of the solids-free wastewater is adjusted with NaOH in a stirred tank. The pH for subsequent precipitation of heavy metals is established in a second stirred tank. Heavy metals are precipitated from the basic wastewater by the addition of sodium sulfide in a stirred reactor. Experience with coal gasification shows that heavy-metal complexes need not be anticipated from gasification in the high-temperature reactor or from gas cleanup.
   The outlet wastewater stream goes to a static mixer, where iron chloride and flocculants are added.
2) Flocculation and sedimentation: This stage of the process consists of a stirred tank for separation of further heavy metals by flocculation and a lamellar classifier for the removal of product solids by sedimentation. The clarifier discharge is pumped to microfiltration.
   The sedimented metal hydroxide slurry is transferred to the static thickener and thence to the recessed-plate filter press. The filter cake mainly consists of metal hydroxide sludge, sulfides, fluorides, and carbonates, and must be disposed of as hazardous waste.
3) Microfiltration: After sedimentation, the finest solids are separated from the wastewater in two alternately operated sand filters. While one filter is in service, the other is being backwashed.
   Filtrate from the downstream collecting tank is used as backwash water. The solids-laden backwash liquor is pumped back to the second stirred reactor of the precipitation stage. Because the fine particles form conglomerates in the sand filter, they are removed by flocculation and sedimentation downstream of the precipitation stage.
4) Reverse osmosis: Two-stage reverse osmosis is used to concentrate the salt solution, which contains mainly chloride and a small amount of sulfates. Part of the first-stage permeate, depending on quality, is recycled to maintain a constant salt load in the second stage. The second-stage permeate is used for cooling, together with the condensate from the evaporation unit. The concentrated salt solution is led to the buffer tank of the evaporator.
5) Evaporation: The solids-free wastewater flows from the buffer tank into the pump suction tank of the evaporation unit, from which it is fed into the forced circulation of the two parallel evaporators. The two evaporator loops each consist of the evaporator proper, the stripper, and the circulation pump.

In the evaporator, the salt solution is heated to its boiling point and part of the liquid evaporates. This portion of the water is separated as vapor in the stripping section. The pump circulates the salt solution until it reaches its final concentration. The solution is then collected in a stirred tank. The forced circulation limits the de-position of salt in the apparatus and piping. An antifoam is added to prevent foaming in the stripping vessel.

After the evaporator, the salt solution is further concentrated in a hydrocyclone and a centrifuge, put into containers, and disposed of externally. The condensate from the evaporator vapor is collected and reused in gas cleaning. The salt-laden wastewaters from the cyclone and centrifuge are recycled to evaporation.

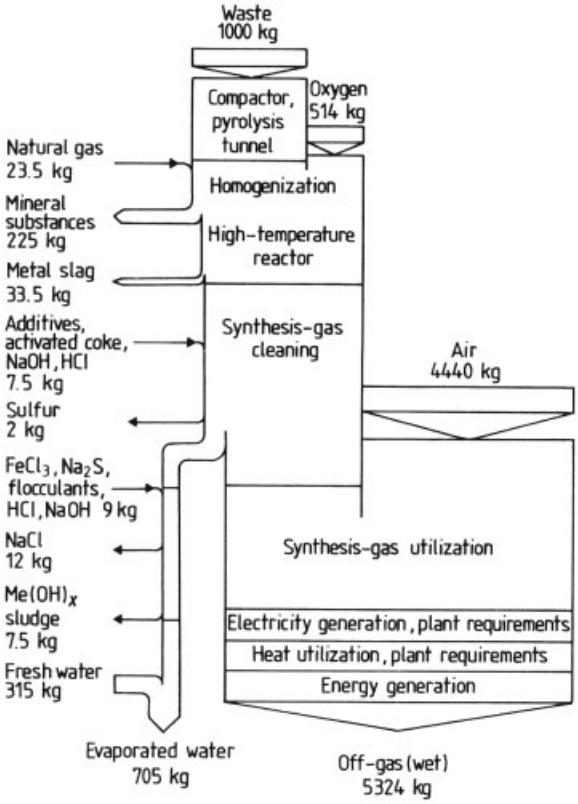

Figure 129. Mass balance of the Thermoselect process

Table 29. Comparison of emission concentrations

	17. BImSchV	Gas engine exhaust	Synthesis gas[a]
O_2, vol %	11 %	11 %	irrelevant
Dust, mg/m^3	10.0	<0.23	0.2
Total C, mg/m^3	10.0	5.0	1
SO_2, mg/m^3	50.0	5.0	<1[b]
HCl, mg/m^3	10.0	0.2	<0.2
HF, mg/m^3	1.0	0.1	<0.1
NO_x, mg/m^3	200.0	10.0	<0.5[c]
CO, mg/m^3	50.0	5.0	
Hg, mg/m^3	0.050	0.020	0.01
Cd, Tl, mg/m^3	0.050	0.002	0.003
total heavy metals, mg/m^3	0.500	0.040	0.02
PCDD/F, ng/m^3 TE	0.1	0.002	0.002

[a] Not comparable with limit values of 17. BImSchV.
[b] Measured as H_2S.
[c] Measured as NH_3.

Emissions and Products. The products of the Thermoselect process are synthesis gas, mineral substances and metal pellets, sulfur, metal hydroxide sludge, and salt from process-water treatment. Emissions depend on the way in which the synthesis gas is utilized.

Synthesis Gas and Gas Engine Exhaust. Most of the energy content of the wastes is used in the production of synthesis gas. The principal components of the synthesis gas are listed in Table 28. The purification of the synthesis gases effectively reduces the levels of contaminants. Table 29 lists the contaminants present in the synthesis gas and the exhaust from the gas engine.

Regardless of the combustion process used (internal combustion engine, industrial heating, steam), the impurity concentrations found in the purified synthesis gas are such that the anticipated emission values are markedly less than the limits mandated by 17. BImSchV.

After cleaning the synthesis gas contains no significant concentrations of substances whose combustion could lead to the formation of polychlorinated hydrocarbon compounds.

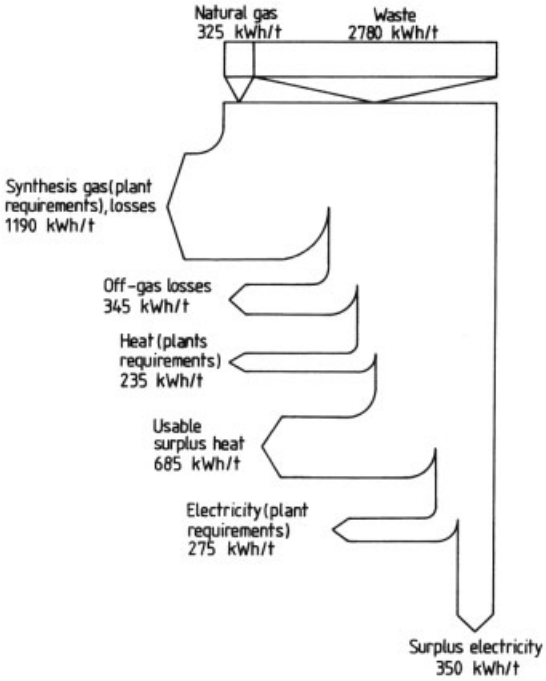

Figure 130. Energy balance of the Thermoselect process

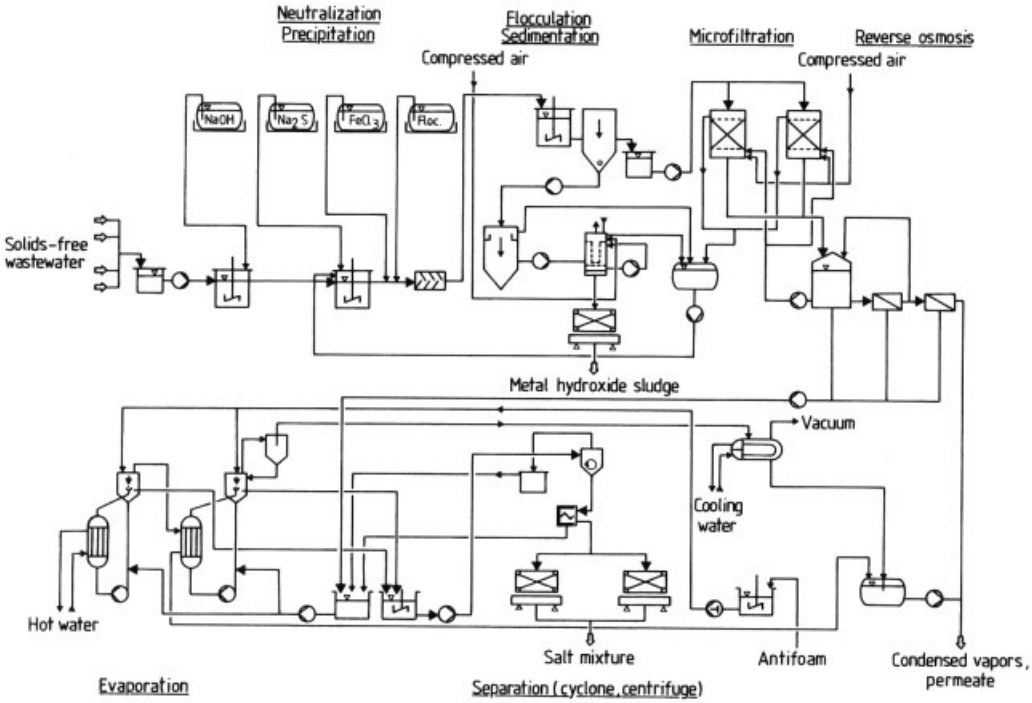

Figure 131. Flowsheet of wastewater cleaning in the Thermoselect process [320]

The content of nitrogen oxides in the off-gas from synthesis gas utilization is essentially controlled by the combustion conditions. The main components, carbon monoxide and hydrogen, lead to an off-gas with low NO_x levels.

Glassy Mineral Substances. The Thermoselect process with optimal energy utilization yields directly usable glassy mineral products with no interruption of the material and energy streams. Suitability of these materials for use as, for example, concrete aggregate has been established in an expert opinion of the Thuringian TÜV [320], [321]. The residues contain sufficient amounts of mineral-forming substances to give eluate-resistant materials.

With regard to the quality of the mineral products, the low residual carbon content of 0.05 – 0.12 wt % and the stable trapping of heavy metal traces are of key importance.

The basalt-like state and the high degree of inertness are responsible for the good elution parameters, as listed in Table 30. The measured contaminant levels in the eluate are low, so that this pelletized product meets the standards of the TA Siedlungsabfall. According to opinions issued by the Rhine-Westphalian and Thuringian TÜV, this material can be deposited or used without special precautions.

The following options exist for the use of the pellet product:

1) In ground form, directly at the construction site (similarly to fly ash from coal-fired plants), as partial replacement for cement in concrete preparation; also as a filler (>90 µm) for asphalt
2) In ground form, with the addition of blowing agents, for the fabrication of expanded granular material with good insulating properties
3) As a blasting medium for metal surface finishing
4) As an aggregate for concrete with partial or complete substitution of one aggregate fraction
5) As a frost-protection layer (subbase material) for highway construction or in bituminous road surfacing
6) For manufacture of mineral fibers directly from the melt

Metal Alloy. The metal produced in the Thermoselect process is low-alloy iron (>90 wt % Fe) with the following principal alloying elements [322]: copper (5 wt %), nickel (0.6 wt %), chromium (0.3 wt %), tin (0.4 wt %), phosphorus (2 wt %).

The concentrations of environmentally relevant elements such as thallium (0.01 wt %), cadmium (<0.001 wt %), lead (<0.08 wt %), mercury (<0.00001 wt %), and arsenic (<0.05 wt %) are low. This metal can be recycled to metallurgical processes, for example as a raw material for copper winning.

Development of the Process. In the 1980s the original idea of the Thermoselect process was compacting wastes to improve landfill utilization. Because the compacted waste packets were unstable, the possibility of thermal surface treatment was investigated [321].

The principal developmental stages of the Thermoselect process are summarized in the following:

1989	Carbonization behavior as a function of pressure, temperature, time; material suitability (10 kg/h)
1990	Carbonization and degassing behavior with compactor, pyrolysis in pressure pipe; testing of melt behavior, instrumentation, and various gas cleaning variants (250 kg/h)
1991	Carbonization and degassing behavior in various pressure- and temperature-stable pusher furnaces up to 20 m long and with cross section of 0.225 m^2, coupled with compactors; testing of gasification behavior, test of lining; burner development, including safety aspects and quick-replacement device; material stability; instrumentation; product quality (1000 kg/h); development of overall concept of full-scale pilot plant and preparation of construction documentation (4200 kg/h)
9/1991	Start of plant construction
12/1991	Completion of building
3/1992	Completion of main installation tests on stable melt discharge, effect of level; tests of safety systems; startup and shutdown conditions; charging control; testing of stability of compactor/tunnel, tunnel/high-temperature reactor, and high-temperature reactor/quench interfaces; continuous operation, reaction to load variations; optimization of gas cleanup, optimization of internal combustion engine operation
6/1992	Sequestration ordered by government authorities
10/1992	Commissioning of entire plant after final installation of evaporator unit; regular plant operation on license basis of 4200 kg/h
1/1993	Start of plant evaluation by Italian, Swiss, and German bodies
4/1993	Dismantling of main modules and inspection by independent bodies to assess stability of individual components and availability
11/1993	Completion of 12-month main testing, including test of a 1.2 MW gas engine
12/1993	Evaluation of experience to define standard Thermoselect plant
4/1994	Approval for continuous operation; restart

Table 30. Comparison of eluate values with the requirements of TA Siedlungsabfall (determined by the German method DEV-S 4 and the Swiss method SET)

Parameter	Classification values		DEV-S 4 [a]	SET [b]
	Landfill class I	Landfill class II		
pH	5.5 – 13.0	5.5 – 13.0	6.5 – 7.9	5 – 5.4
Conductivity, µS/cm	≤10 000	≤50 000	13.0 – 29.0	94 – 262
AOX, µg/L	≤0.3	≤1.5	–	–
TOC, mg/L	≤20	≤100	0.01 – 0.03	0.01
Phenols, mg/L	≤0.2	≤50	0.03	0.01
CN^-, mg/L	≤0.1	≤0.5	0.02	0.013 – 0.021
F^-, mg/L	≤5	≤25	0.02 – 0.04	0.02 – 0.04
NH_4-N, mg/L	≤4	≤200	0.03 – 0.1	0.04 – 0.5
As, mg/L	≤0.2	≤0.5	0.1	0.01
Pb, mg/L	≤0.2	≤1	0.01	0.01
Cd, mg/L	≤0.05	≤0.1	0.001	0.001
Cr(VI), mg/L	≤0.05	≤0.1	0.02	0.02
Cu, mg/L	≤1	≤5	0.015 – 0.085	0.012 – 0.2
Ni, mg/L	≤0.2	≤1	0.005	0.005 – 0.18
Hg, mg/L	≤0.005	≤0.02	0.0005	0.0005
Zn, mg/L	≤2	≤5	0.005 – 0.013	0.094 – 0.23
SO_4^{2-}, mg/L			0.15 – 2.53	1.42 – 1.77

[a] DEV-S 4 = Deutsches Einheitsverfahren S 4. [b] SET = Schweizer Eluat-Test.

Several aspects of plant operation, especially emissions and product quality, and of plant safety were assessed by independent experts. The safety of the pilot plant, in particular with regard to fire and explosion protection, was determined by Italian licensing authorities.

Tests in the large pilot facility led to optimization of the following points:

1) Improved cooling capacity in the quench water system
2) Use of an improved lamellar classifier system and discharge of solids from the quench loop for recycle to thermal treatment
3) Integration of direct sulfur removal, including regeneration of the Fe(II)/Fe(III) complex
4) Integration of an internal combustion engine/generator system for power generation

The capacity limitation of 4.2 t/h specified for the erection of the Thermoselect plant will remain in place because all engineering equipment selections and the design of the gas and water cleanup units, including evaporation, were based on this figure for approval-related reasons.

9.4.3.3. Noell Conversion Process

Description of the Process. The Noell conversion process combines thermal conditioning (pyrolysis) with an entrained-bed gasification stage. The heterogeneous starting materials (e.g., residues from treatment of household and hazardous waste, electronics scrap, and automobile scrap processing) are homogenized into coke in the pyrolysis step, then ground. These intermediate products are inlet to entrained-bed gasification together with the pyrolysis gas and the collected condensates. Residues in dust form, such as dried and ground sewage sludges, pulverized plastics, waste oils, pumpable sludges and slurries, paint residues, and contaminated liquids, can be fed directly to the entrained-bed gasifier. Figure 132 illustrates the design of the Noell conversion process.

Waste Pretreatment. After size reduction in rotary shears to <50 mm, the wastes are either inlet directly to pyrolysis or, depending on their moisture content, to a preliminary drying stage. The purpose of this step is to improve the thermal efficiency, especially where water contents are high. The waste heat from the process is used.

Pyrolysis. The pyrolysis stage is based on experience with the hazardous-waste pyrolysis plant at Salzgitter, Germany, in which a rotary kiln with a diameter of 2.80 m and a length of 28 m was tested. Figure 133 shows the construction of the pyrolysis stage.

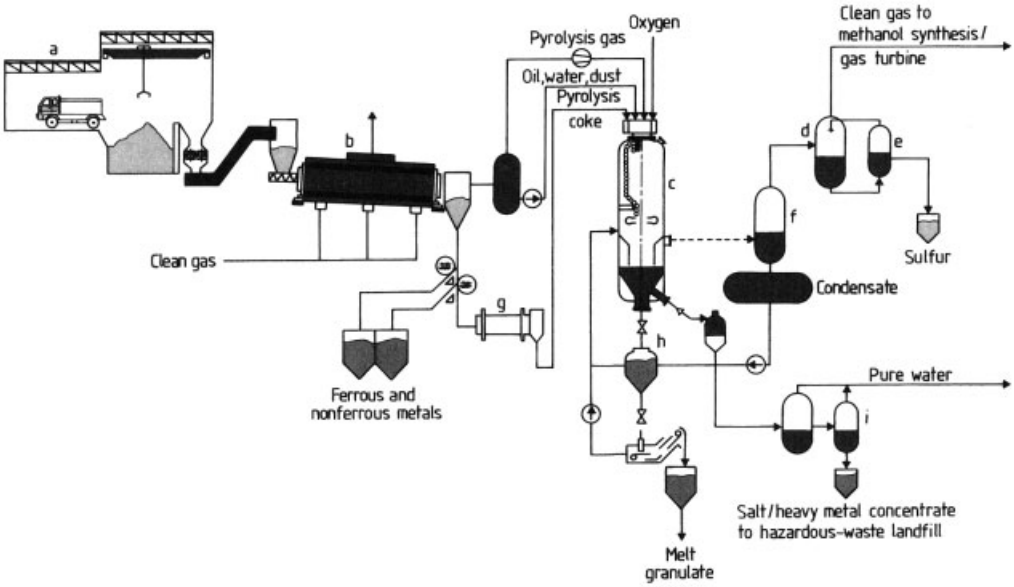

Figure 132. Flowsheet of the Noell conversion process [323]
a) Bunker, size reduction; b) Pyrolysis; c) Entrained-bed gasification; d) Scrubber; e) Regenerator; f) Cooler; g) Mill; h) Lock

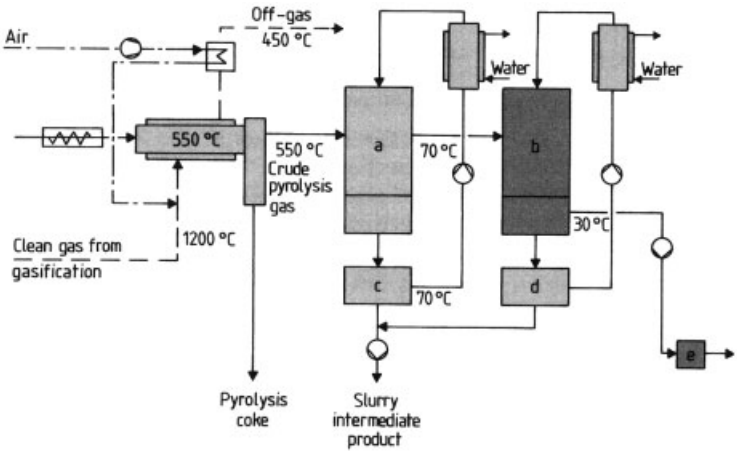

Figure 133. Flowsheet of the pyrolysis section of the Noell conversion process [323]
a) Quench cooler; b) Final cooler; c) Decanter 1; d) Decanter 2; e) Gas tank

Residual waste is discharged through a lock system, blanketed with nitrogen to exclude oxygen. A pusher screw below the lock system compacts the waste and charges it to the pyrolyzer. The compacted plug prevents the escape of pyrolysis gas through the charging system.

In the rotary kiln, the waste is heated to a temperature between 650 and 750 °C under air exclusion. The waste is pyrolyzed, yielding crude pyrolysis gas and pyrolysis coke as intermediate products; these are forwarded to intermediate treatment.

The pyrolysis unit is indirectly heated with purified synthesis gas from the gas cleaning process. Heating the pyrolysis kiln requires 20–25 % of the clean gas produced. The use of low-temperature burners to limit NO_x formation makes it possible to meet the standards of the 17th Federal Pollution Control Regulation (17. BImSchV).

After separation of the pyrolysis coke, the crude pyrolysis gas discharged from the coke separator at a maximum temperature of 500 °C is cooled to 70 °C against cooled recirculating product. The various liquid hydrocarbons, as well as water and coke dust if present, are eliminated here. The separation of a solids-poor fraction for recycling takes place in a downstream decanter. In a second cooling stage working on a similar principle, the pyrolysis gas is brought to 30 °C before going to gas storage and compression. Utilization of pyrolysis gas in the entrained-bed gasifier increases the flexibility of the system to handle strongly fluctuating ash levels. At the same time, hydrocarbons, and particularly also chlorinated hydrocarbons, in the synthesis gas components are decomposed. The slurry product from the two decanters is also inlet to the entrained-bed reactor.

Treatment of Pyrolysis Products. The pyrolysis stage is followed by separation of pyrolysis coke, pyrolysis gas, and pyrolysis vapors and treatment of these products for use in the entrained-bed gasifier.

First, the pyrolysis coke is cooled. Next, ferrous and nonferrous metals are removed by magnetic and eddy current separators, and the coke is ground to <0.5 mm. The separation of inert materials is not carried out. The ground pyrolysis coke is delivered to the gasification burner by a pneumatic conveyor.

The gas and vapor products from pyrolysis are quenched, yielding an oil – water – dust slurry that contains the readily condensable species. The gas is then indirectly cooled to separate water and condensate. The slurry and the pyrolysis gas are separately forwarded to the gasification reactor by pumps and compressors.

Because pyrolysis and gasification take place in separate apparatus, the two substages can operate independently if intermediate storage is provided for the products.

Entrained-Bed Gasification. The intermediate products are fed to the entrained-bed gasifier, where they are converted to crude synthesis gas with technical-grade oxygen. The gasifier consists of a cylindrical reaction chamber bounded by a water-cooled pipe wall. The gasification burner is located at the top of the reaction space and delivers the material for gasification and the oxygen, and maintains a pilot flame generated by a combustible gas of external origin (natural gas, digester gas, pyrolysis gas) or recycled clean gas. Ignition and monitoring devices are integrated in the burner (Fig. 134).

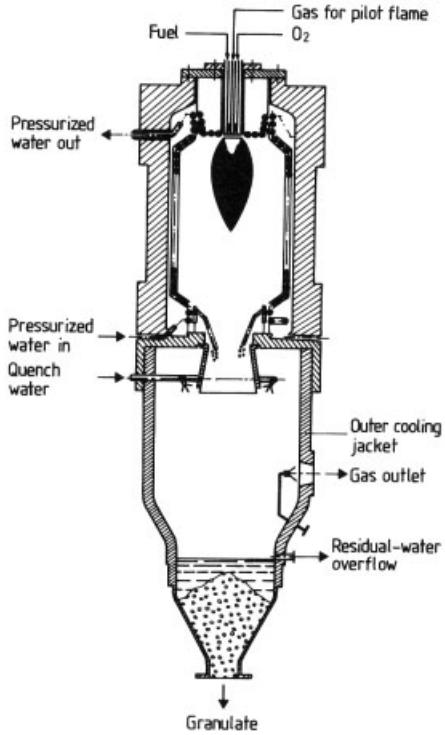

Figure 134. Entrained-bed gasifier used in the Noell conversion process [323]

The gasification feed, in dust form, is delivered by a pneumatic metering system consisting of a supply bin and pressure locks operated in alternation. Inert gas is injected to bring the material to the plant pressure of 25 bar. The material then flows by gravity to a special metering vessel. Injection of an inert conveying gas into the bottom of the metering vessel fluidizes the dust in a localized manner, and in this state (dense suspension in conveying gas) it is delivered to the gasification reactor burner by means of a pressure difference.

In the reactor, the substoichiometric quantity of oxygen and the carbon from the organic feed materials are converted to CO_2 in a flame reaction. The CO_2 is then reduced to CO according to the Boudouard reaction. In a parallel process, water is cleaved down by the water-gas reaction at final temperatures of 1400 – 1700 °C, so that a crude synthesis gas, rich in CO and H_2 and

free of hydrocarbons, is produced. This strongly reducing gas is led downward into the quench.

Mineral constituents of the feed material form a slag film, which flows down into the quench space via the reaction space wall, which is protected by a thin layer of a special rammed mass and solidified slag. The gas, saturated with water vapor, exits the quencher at the side, while the granulated product is collected in the water bath and discharged through a feeder.

The quantity of synthesis gas produced is 0.7 – 1 m^3/kg of waste and sewage sludge, depending on the water and ash contents in the feed. The purity of the technical oxygen can be adapted to the intended use. The oxygen ratio is chosen between 0.3 and 0.6 so that the exit temperatures of the crude gas ensure continuous flow of the glassy melt without supplementary heating.

Synthesis Gas Cleaning. The crude gas, saturated to the dewpoint after quenching, is next subjected to the following purification stages:

1) Venturi scrubber to remove dust and salt
2) Condensation of water vapor
3) Desulfurization (removal of H_2S and COS)
4) Optional processing to synthesis gas for methanol synthesis (Fig. 135)

To remove dust and salt, the crude synthesis gas is passed through two Venturi scrubbers in series. Metal oxides can also be precipitated as metal sulfides in this stage by using hydrogen sulfide. The wastewater from these stages is fed to quenching.

At the plant pressure of 25 bar and a saturation temperature of 205 °C, the crude synthesis gas has a steam content of 65 %, so that a low-pressure steam (about 6 bar) for district heating or internal use can be obtained by cooling. Condensation of the water vapor when the gas is cooled also removes any remaining traces of salt and fine solids. This is a prerequisite for the downstream desulfurization step.

Sulfur is present as hydrogen sulfide (ca. 90 %) and carbonyl sulfide (ca. 10 %). Oxidation processes involving iron chelate solutions or adsorption processes involving activated coke or metal oxides are employed to remove this sulfur. After this stage, the clean synthesis gas is suitable for use as a feedstock or fuel. Depending on the use, the gas may be treated in a variety of ways, for example by crude gas converting to establish a desired H_2/CO ratio in a feedstock gas.

By recycling the quench and scrub liquors and the condensate produced by gas cooling, the quantity of process wastewater requiring purification can be restricted to the amount entering with the residues and wastes. The process can thus be operated without generating wastewater.

Crude Gas Scrubbing. Like the use of synthesis gas as a feed to catalytic synthesis processes, its use to generate electricity, whether in gas turbines or internal combustion engines, requires the complete removal of fine dust, volatile metal oxides, and salt mists. The desulfurization stage downstream of the crude gas scrubbing can tolerate only limited concentrations of these contaminants. If gasification is performed under pressure, the use of Venturi scrubbers followed by a heat-recovery boiler (Fig. 135) offers a fairly simple solution. When large amounts of water vapor are condensed from the crude synthesis gas in the heat-recovery step, even small aerosol particles are trapped in the condensates, the larger particles having been removed in the Venturi scrubbers. The scrubbing liquor also dissolves part of the hydrogen sulfide. This is a prerequisite for the precipitation of volatile heavy metal oxides as heavy metal sulfides, which are insoluble and not classified as toxic.

Desulfurization. The gasification of residues and wastes under pressure makes it possible to remove sulfur down to the level of 1 – 10 mg/m^3. This is an order of magnitude better than the 50 mg SO_2/m^3 limit set by the 17th Federal Pollution Control Regulation. What is more, the volume of the gas stream is much smaller. These gains are possible only if carbonyl sulfide is eliminated by catalytic hydrolysis. The removal of hydrogen sulfide on activated coke or carbon is a simple technology and one proven over decades. If the stoichiometric quantity of oxygen for oxidizing H_2S to elemental sulfur is provided, the oxidation process goes to completion on the internal surface of the coke, and elemental sulfur is deposited there at the same time. It is desirable to use a moving bed in the adsorber, with fresh coke inlet at the head and the crude gas led in countercurrent. The spent activated coke is withdrawn at the bottom and recovered.

In the Sulferox process, hydrogen sulfide is oxidized by Fe^{3+}, dissolved as an iron chelate. The crude gas flows through a bubble-column

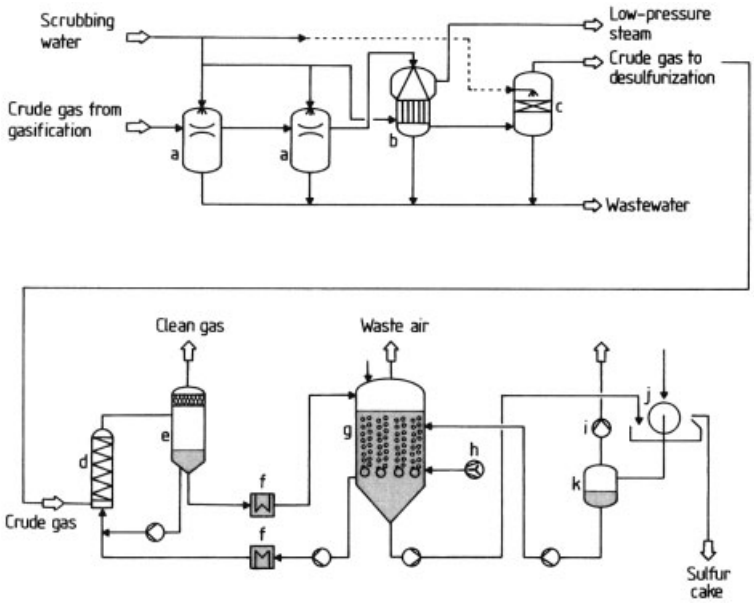

Figure 135. Flowsheet of synthesis gas purification in the Noell conversion process
a) Venturi scrubbers; b) Waste-heat boiler (partial condensation); c) Separator (mist eliminator); d) Absorber; e) Separator; f) Startup heaters; g) Regenerator; h) Air blower; i) Vacuum pump; j) Drum filter; k) Filtrate tank

absorber before leaving the downstream separator as cleaned synthesis gas. The iron reduced to Fe^{2+} by oxidation of hydrogen sulfide is passed to the regenerator and reoxidized to Fe^{3+} with air. It can then be recycled to the absorber. Elemental sulfur is separated in a drum filter and washed.

Wastewater Treatment. The type of process wastewaters inlet to treatment depends on the design of the circulation streams (see Fig. 136):

Stream 1:	Wastewater from granulate product discharge (ash removal)
Stream 2:	Venturi scrubbing liquor
Stream 3:	Excess water from quencher

For the utilization of heat internal to the process, the hot (205 °C) scrubbing and quench liquors are first used in the evaporator. After the water from granulate discharge is added to ensure optimal closed-circuit operation and minimize the quantity inlet to the wastewater treatment system proper, solids are mechanically removed by a hydrocyclone in a first process step. The solids-rich stream from this stage contains essentially slag fines. It is dewatered in a belt-type filter press and then fed back to the gasification unit to be incorporated in the granulated product. In the second process step, milk of lime is added to precipitate sparingly soluble salts. Any heavy metals still in solution at this point are collected by a selective precipitating agent.

The precipitated salts, together with solids not removed in the hydrocyclone and soot particles, coagulate in the flocculation stage and are removed by gravity in the lamellar classifier. The product slurry is thickened, flocculant is added, and the slurry is dewatered to >35 % dry solids in a recessed-plate filter press. After precipitation/flocculation, the wastewater is in such a condition that much of it can be reused in the gasification process. Wastewater not returned to the process is subjected to two-stage cyanide oxidation, in which hydrogen peroxide is used as the oxidizing agent. Next, controlled amounts of hydrochloric acid are added to set an acid pH, so that ammonium is retained in the concentrate in the subsequent evaporation step. The last stage in the process is evaporation of the wastewater and drying of the product concentrate.

Because the hot stream of wastewater contains sufficient energy, evaporation is designed as a one-stage process. This stage consists of a vapor separator, a forced-circulation pump,

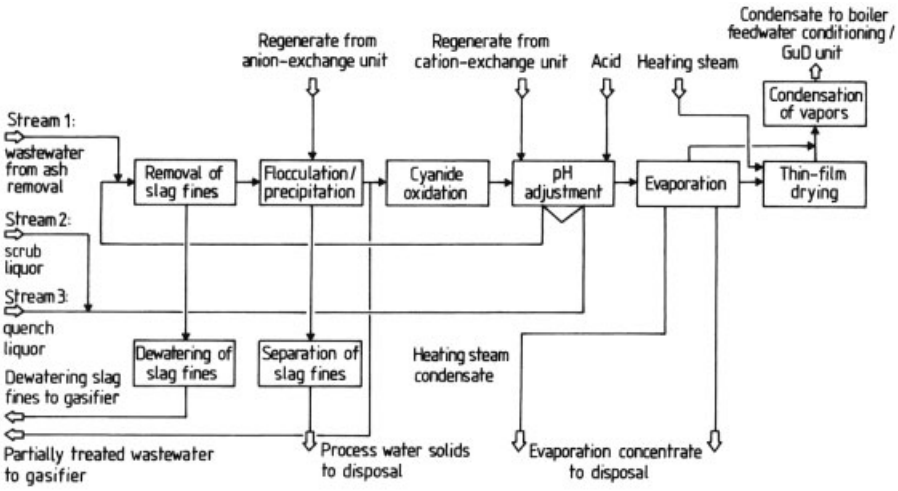

Figure 136. Block diagram of wastewater treatment in the Noell conversion process [323]

and a round-pipe heat exchanger. The evaporation temperature is chosen so that the flash steam obtained from the 205 °C wastewater provides the entire heat requirement for the evaporator. Because the concentrations and quantities are such that the wastewater would otherwise be too strongly concentrated, it is diluted by the return of a substream of distillate. A thin-film dryer heated with superheated steam (max. 240 °C) follows, in which the concentrate is dried to >90 % dry solids and packaged for landfill disposal. This mixed salt can be converted to technical-grade products, but the cost is high and this use is therefore not recommended.

The vapors from the second stage of the evaporation unit and thin-film dryer are condensed against water in surface condensers. All the vapor condensate is then pumped into its own tank, from which it is withdrawn for feedwater conditioning and preparation of chemicals.

Emissions and Products. *Synthesis Gas.* The content of $CO + H_2$ in the synthesis gas lies between 89 % for contaminated waste oil with a high heating value and 67.1 % for straw as a biofuel. Sewage sludges and residue fractions from residential waste fall in between.

Methane, the most stable hydrocarbon, is virtually absent. Higher hydrocarbons are only present in nano- or microgram traces. The concentration of nitrogen stems from the organic matter as well as the inert gas used in pneumatic conveying, especially in the case of residential waste fractions and straw. The nitrogen content can be significantly reduced if gas from the process is recycled.

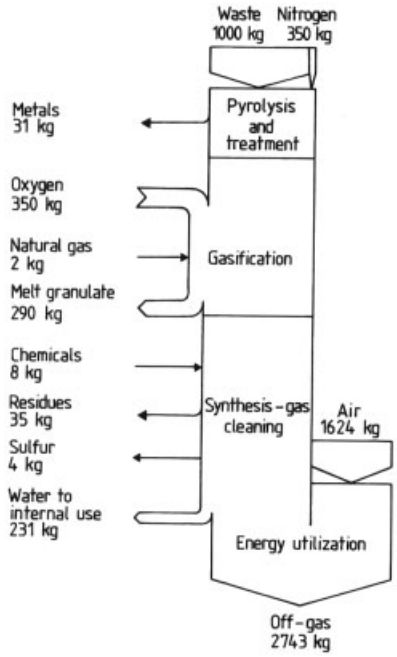

Figure 137. Mass balance of the Noell conversion process

Sulfur entering with wastes is converted 90 % to hydrogen sulfide and 10 % to COS; the ratio does not depend on the type of fuel. Part of the

hydrogen sulfide is dissolved in the quench and scrub liquors, so that heavy metals not trapped in the granulate product are converted to insoluble heavy metal sulfides and precipitated. Ammonia and hydrogen cyanide are present at lower concentrations, and are removed from the gas and dissolved in the scrub liquor. They are eliminated in wastewater treatment. The clean gas has a heating value of 8 – 12 MJ/m^3 and can be used as a feedstock (synthesis gas) or fuel. After sulfur removal, the gas surpasses the standards of the 17th Federal Pollution Control Regulation.

Figures 137 and 138 show the mass and energy balances.

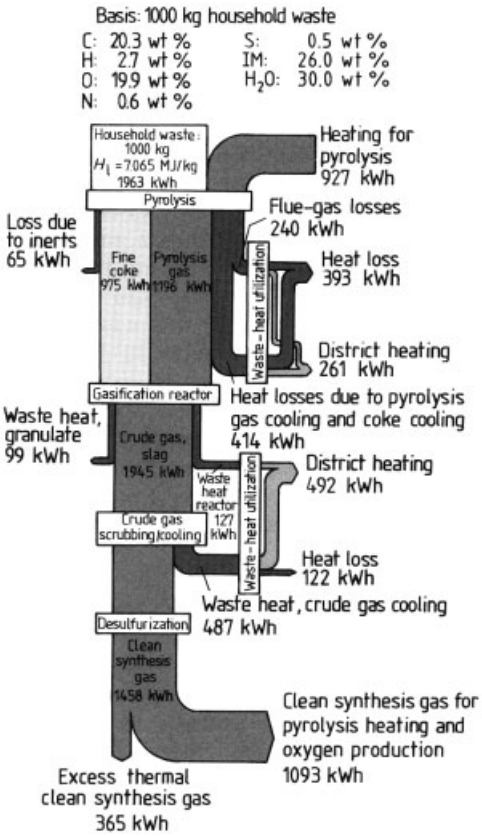

Figure 138. Energy balance of the Noell conversion process
IM = Inorganic material

Melt Granulate. Inorganic species inlet with the wastes are discharged from the gasifier as a melt and pelletized in the quench chamber. The granulate represents a sink for nonvolatile heavy metals. A small fraction is trapped in the granulated product. Most is precipitated in sulfide form and filtered out of the wastewater as solids. The process water solids are the sink for volatile heavy metals. A key point for economic utilization of the pellet product is its elution behavior. Elution values for pelletized melt product from residential waste, bituminous coal firing, and Freiberg gneiss (a native rock) are similar.

Solids from Wastewater. Solids from wastewater, consisting of process water solids and evaporator residue, must be discharged from the plant. According to the wastewater treatment flowsheet, process water solids in the form of filter cakes are obtained after flocculation and precipitation of heavy metals and before the partly treated water is forwarded to other operations.

In general, an overall technology that discharges no wastewater is needed. The excess water is evaporated by waste heat. The evaporation residue is a mixed salt, which is classified as nontoxic and can be disposed of in hazardous-waste landfills. The composition and balance of the solid residues are as follows:

Silicon	23 – 46 wt %
Aluminum	9 – 16 wt %
Calcium	13.5 – 39 wt %
Sodium	0.5 – 0.9 wt %
Magnesium	2.1 – 3.1 wt %
Potassium	0.4 – 1.6 wt %
Iron	6 – 19 wt %
Copper	0.01 – 0.12 wt %
Chromium	0.006 – 0.044 wt %
Nickel	40 – 180 mg/kg
Zinc	0.004 – 0.011 wt %
Lead	0 – 0.01 wt %
Cadmium	0 – 0.9 mg/kg
Mercury	0 – 0.6 mg/kg
Chlorine	0
Phosphorus	4 – 17.5 wt %
Sulfur	0 – 0.1 wt %
Carbon	0 – 0.3 wt %

The quantities of process water solids and evaporation residue, which are approximately equal, are governed by the concentrations inlet with wastes and residues. These levels of lead and zinc mean that the process water solids are a zinc – lead blende that can be processed by metallurgical methods.

Development of the Process. Entrained-bed gasification was developed independently and concurrently by Noell, Würzburg, and the Deutsche Brennstoffinstitut, Freiberg. Scaleup

began around the end of the 1970s and a full-scale experimental plant was operated from 1984 to 1988.

A 3 MW demonstration plant is available at Freiberg for fundamental research, especially with substances not covered by previous experience. This has been in operation since 1979. Some 60 substances have been gasified and their material and energy balances worked out.

An entrained-bed gasifier with a capacity of 130 MW, corresponding to about 30 t/h of feed, has been producing town gas on an industrial scale since 1988. The feed consists of condensate products from fixed-bed pressure gasification of lignite. This plant, along with the fixed-bed pressure gasifier, is now part of the waste recovery center of the Lausitzer Bergbau-Verwaltungsgesellschaft.

Pyrolysis was tested in a pilot facility at Salzgitter. This plant was operated for half a year in 1985. After renovation, it was on stream for two years starting in 1986, pyrolyzing 10 396 t of waste in 6151 hours.

After assessment of extensive tests with the full-scale pyrolysis system at Salzgitter, approval for continued operation was granted, and the start of disposal operation is set for late 1995.

10. Deposition

10.1. Deposition Aboveground

10.1.1. Introduction

In this section, the aboveground disposal of wastes is referred to as landfill disposal or landfill.

In Germany as in many other countries of the European Union, landfill disposal has been accorded a low priority in official waste-management concepts. In practice, however, the majority of wastes continue to be deposited in landfills. The term landfill disposal includes a broad spectrum of disposal qualities, from unmonitored "dumps" with no technical installations up to well-engineered, equipped, and monitored landfills. This section, however, discusses only landfills where pollution of the environment can be largely eliminated.

10.1.2. Classification of Landfills

In most countries, wastes are placed in special kinds of landfills according to their demonstrated or presumed environmental impacts. The classification schemes often have not been precisely defined. An exception is the German federal state of North Rhine-Westphalia (NRW), which in 1987 stated exact definitions of six classes of landfills [331]:

1) Landfill class 1: Soil disposal
2) Landfill class 2: Disposal of mineral products (e.g., demolition waste)
3) Landfill class 3: Disposal of municipal solid waste (MSW)
4) Landfill class 4: Disposal of commercial and industrial wastes
5) Landfill class 5: Disposal of hazardous wastes
6) Landfill class 6: Underground disposal of hazardous wastes

Along with the general information required for classifying wastes under this scheme, disposal was also made dependent on the contents of specific substances in the waste or in the leachate.

Since 1993, the assignment of wastes to disposal sites in the Federal Republic of Germany has been governed by the "Technical Directive on Municipal Solid Wastes" (TA Siedlungsabfall, with two landfill classes) [332] or, for hazardous wastes, TA Abfall [333] with one class. These two regulations also limit the concentrations of certain substances in waste and leachate. Although the classification system is largely the same as that cited above, there is a crucial difference: in the NRW system, solid wastes can always be assigned to a landfill class, but the constraints on the classes set up by the TA Siedlungsabfall and TA Abfall are so narrow that in many instances it is necessary to treat the wastes before they can satisfy the classification criteria. The purpose of these new regulations is to secure the pretreatment of wastes such that an environmentally neutral form of disposal is ensured. Exceptions are built into TA Siedlungsabfall in the form of dedicated disposal sites for wastes not complying in all respects with the landfill class requirements.

Insofar as classification systems do exist in the several countries, they correspond to one of the two systems just discussed (sometimes in modified form).

A special class of disposal site exists in Great Britain and Ireland: the "co-disposal" landfill. This type of site handles MSW as well as industrial and hazardous wastes all together. The basic concept of this type of landfill is to accept the last-named wastes only in such quantities that the emissions by the leachate path are no higher than those of an MSW landfill. Tests are performed to determine the levels of various contaminants that can be trapped or degraded in the MSW. The result provides a basis for establishing total and short-term limits (per week or month) on the amounts of each contaminant that can be deposited in the landfill as a function of the amount of MSW deposited there. The maximum acceptable quantities of industrial and hazardous wastes are then derived from these contaminant limits and the concentration of the substances in the wastes. Wastes should be deposited in such a way that the contaminants escaping from them can be safely taken up by the municipal solid wastes and either immobilized or degraded.

10.1.3. Landfill Siting

The EU countries have no exact quality criteria for the selection of landfill sites.

The most important criteria for siting in most countries are geological and hydrological constraints. For example, the soil should have a low permeability and a high potential for retaining contaminants. The groundwater at the planned landfill site should not be used in producing drinking water either currently or in the future. These requirements, also referred to as geological barriers in Germany (see Section 10.1.4), represent the key requirements for a disposal site.

In addition, a number of special criteria must be taken into account in landfill siting: natural conservation and landscape protection regions, bodies of water, inhabited areas, infrastructure, and so forth. Sites where there is a danger of earthquake or mining subsidence must also be eliminated as far as possible.

While the last-named criteria can be researched in existing documents, in Germany a detailed exploration of the geological barriers at potential sites is required. Once this exploration has been completed for the prospective sites, a comparison should be performed as to environmental and geographical compatibility, and the most suitable site is selected. Then, by means of detailed investigation of the selected site, it must be demonstrated that the first explorations of the geological barriers and analyses of soil stability are correct in all points or that the site actually surpasses the first findings.

In order that landfills can be constructed even in regions where the geological conditions are poor, TA Siedlungsabfall permits the use of technical measures to supplement (or, in some German states, to replace entirely) inadequate geological barriers.

10.1.4. Landfill Safety Analyses

Recurring problems with landfills and the resulting impacts on the environment have made it necessary to conduct thorough environmental safety analyses for landfills. One consequence of these stringent safety analyses in Germany is the multibarrier concept. In contrast to disposal sites in the past, where a bottom liner was virtually the only safeguard against environmental impacts of the landfill, the multibarrier concept requires the use of a number of independent emission barriers [334], [335]. The following kinds of barriers are contemplated:

Barrier 1 is the pretreatment of waste in order to immobilize contaminants, since the long-term mobility of waste constituents is one of the most important parameters relating to the hazard potential of a disposal site.

Barrier 2 is the geology and hydrology of a landfill site. Because of the extremely long service life of a landfill, it cannot be excluded that artificial barrier systems may partially or completely fail over time. If this happens, the geological and hydrological barrier must ensure that escaping contaminants migrate slowly and at low concentrations.

The landfill material with its forecast behavior represents barrier 3. The processes taking place in the body of a landfill must be largely describable and predictable. The stability of the

deposited material and the immobility of its constituents thus constitute barrier 3.

The landfill foundation seal or bottom liner with leachate collection system is designated as barrier 4. The purpose of this system is to prevent leachate, possibly heavily polluted, from escaping into the surroundings. The liner and drainage system must remain fully functional for as long a time as possible.

The top liner with pecipitation runoff constitutes barrier 5. This system must control the penetration of precipitation into the landfill material and thus reduce its effects on barriers 3 and 4. At the same time, the deposited waste must be protected against weathering for as long as possible.

Barrier 6 comprises the maintenance, monitoring, and repair of the barrier systems. In view of the long existence of a landfill and the limited service lives of the technical systems, this barrier takes on special importance. Monitoring makes it possible to keep track of the functionality of the barrier systems, while repairability means that functionality is drastically extended over time.

Only the future will tell, however, how effective and feasible such a multibarrier system will be.

10.1.5. Lining and Drainage Systems

Since about 1980, in response to the finding that many unlined landfills had caused serious groundwater contamination, it has been increasingly recognized that the bases of landfills must be as tightly sealed as possible.

At first, natural mineral foundations were used, but also clay and plastic liners. Although natural inorganic materials were found in laboratory tests to have the required low permeabilities, in many cases the materials turned out to have a markedly higher permeability in natural deposits.

Mineral liners consist mainly of clay and other fine components, compacted in such a way that the individual layers exhibit extremely low permeabilities ($k < 5 \times 10^{-10}$ m/s). This type of liner is similar in makeup and structure to natural unconsolidated material and thus shares its long-term stability, but substances such as concentrated acids and alkalies can attack the structure and thereby increase the permeability.

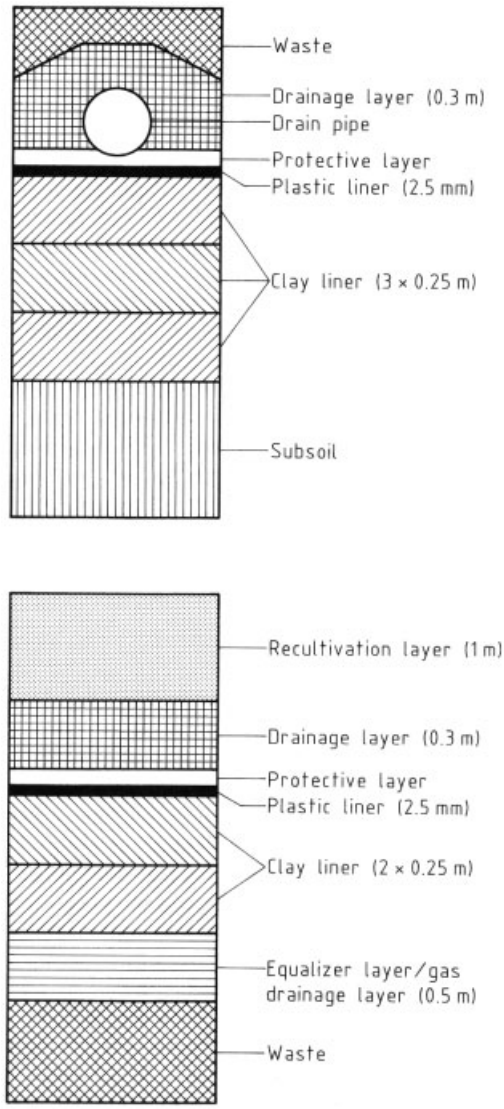

Figure 139. Examples of a composite bottom liner (A) and a top liner (B) [332]

The plastic material most commonly used for liners is high-density polyethylene (HDPE) film. This plastic liner material has the drawback that its long-term stability (>50 years) is unknown. In contrast to clay liners, plastic liners are almost absolutely watertight, provided that the welds between sections are tight. Some solvents, however, permeate relatively fast through plastic liners without damaging them.

In order to combine the advantages and offset the disadvantages, the TA Siedlungsabfall prescribes a composite seal system for use in Germany [332]. The composite is made up of a mineral component and a plastic liner (see Fig. 139 A). The two liners must be placed contiguous to one another.

A liner system alone cannot, however, safeguard the soil against environmental impacts, because the leachate retained at the base of the landfill would escape through even the smallest flaws. Therefore, leachate reaching the liner must be removed as promptly as possible. Experience has shown that effective drainage is possible only with a fairly permeable gravel layer (ca. 0.3 – 0.5 m thick) with supplementary drainage pipes (see Fig. 139). It has proved necessary to use extra-coarse material (grain size 16 – 32 mm) for the drainage layer, since the leachate generated from many wastes contains substances that could precipitate outside the waste fill (e.g., in the drainage layer). When municipal solid wastes are disposed of, it is often required that they be biologically pretreated before or at the beginning of deposition so that reactions in the waste fill will not liberate acids that would dissolve and transport such products. The risk of encrustation decreases with increasing volume of the drainage layer and with increasing pore size. At present there is no method for removing encrusted material from the drainage layer. Because such deposits also threaten the drainage pipes, the pipes are regularly inspected with video cameras, then flushed with a high-pressure (8 MPa) water jet if necessary. Annual inspection and flushing has proved quite beneficial in terms of the functioning of the drainage system. These practices have restricted the length of drain pipes to ca. 300 m. Because any structural foundation (e.g., shafts) placed on the foundation liner must be regarded critically, the maximum pipe length has often dictated the maximum size of the landfill. To avoid this restriction, landfill designs in Germany increasingly use galleries on the landfill base, with leachate pipes (and sometimes also gas collection pipes) running into the galleries. The pipes must be designed so that the drainage pipes can be inspected and flushed at any time without admitting air from the outside. In MSW landfills, because of landfill gas diffusing into these galleries (see Section 10.1.6.5), special explosion protection measures are required.

Even when the pipes end in shafts outside the landfill, care must be taken not to admit ambient air into the pipes and the drainage layer, because air often promotes the precipitation of inorganic compounds from the leachate and biogenic deposits. At the same time, inspection and maintenance must be possible at all times.

To prevent the formation of leachate from landfills over the long term, TA Abfall and TA Siedlungsabfall mandate a "final cover" after closure of the landfill. Figure 139 illustrates such a cover. An equalizer layer must be located underneath the top liner proper to allow the collection of any gas that forms. Above the liner is a layer of soil in which vegetation can grow; this layer must be carefully drained so that precipitation water does not accumulate on top of the liner. Observation has shown that the quantity of leachate is drastically cut by the final cover, but it is not possible to say now whether complete avoidance of leachate is possible. A final cover can perform its intended function only if it is monitored and maintained over a long period (centuries or millennia).

10.1.6. Leachate and Gas Formation in Landfills

10.1.6.1. Processes in the Landfill

A variety of processes take place in all landfills, depending on the type of wastes. These processes are biological, chemical, and physical in nature, or a combination of any or all of these. In only a few types of landfill, such as those for demolition waste and soil disposal, can it be assumed that these processes will be of about the same order of magnitude as processes in natural rock and soil and thus of secondary importance. The leachate formed by precipitation dissolves substances from the wastes and also influences the various processes, generally accelerating them. The processes occurring in landfills are extremely complex and cannot be described in detail.

Some examples of individual processes that can occur in a landfill, depending on the nature of the wastes as well as the ambient conditions in the waste fill, are the following [336]:

1) Precipitation of metals as hydroxides, oxides, carbonates, sulfides, silicates, humates, etc.
2) Mobilization of metals by formation of complexes
3) Mobilization of metals as a result of the lower pH in the acetic phase (see below) or when the wastes contain acidic inorganic substances
4) Changes in solubility behavior due to aging

Except in dedicated landfills with wastes of simple composition, there is always a very complicated interaction of these processes, which ultimately determine the nature of the emissions.

The complexity of the processes becomes much greater if the wastes have organic constituents. Dissolved organic substances take part in some of the processes, and all organic materials (plastics being among the few exceptions) are subject to biological reactions, which vary in type and intensity.

These biological processes have been extensively studied in MSW landfills [337–343]. Immediately after deposition and at the outer surface of the landfill (to a depth of ca. 1 – 1.5 m), aerobic reactions take place. The microorganisms involved require oxygen to degrade organic matter, generating CO_2, H_2O, biomass, and heat as end products. Generally, wastes are exposed to atmospheric oxygen for only a relatively short time, and this means that only the readily degradable substances react. As soon as the oxygen supply ends, anaerobic conditions come into being; these lead to methane fermentation, which continues for a long period (methanogenic phase). During this time, organic substances are converted to CH_4 and CO_2 as well as small amounts of biomass and energy. This anaerobic process proceeds in three phases that can be relatively exactly distinguished. For the reaction mentioned above, these phases must be in equilibrium. The three steps in the anaerobic process are as follows:

1) Hydrolysis: facultative bacteria convert the organic substrate (proteins, fats, carbohydrates, etc.) into simple organic substances such as amino acids, glucose, etc.
2) Acetic phase: the end products of hydrolysis are converted by anaerobic bacteria to volatile fatty acids, CO_2, and H_2.
3) Methanogenic phase: acetogenic bacteria convert volatile fatty acids to acetic acid, CO_2, and H_2; methanogens immediately convert these end products to CH_4 and CO_2. Both groups of bacteria are extremely sensitive, strictly anaerobic, and substrate-specific.

All three phases are collected under the term methanogenic phase. In industrial application this process is held at equilibrium by means of the operating conditions and monitoring. Because such measurement and control are not possible in a landfill, equilibrium is not attained, at least in the first phase of waste deposition. As a result, the end products of hydrolysis and acetic fermentation, which are formed quickly, are transported away with leachate through the fill; in this way they either alter the milieu for processes taking place elsewhere or they become contaminants in the leachate reaching the leachate collection system. Equilibrium is established only when the slow-growing methanogenic bacteria develop, mediating the reaction to the gases CH_4 and CO_2 and thus reducing the organic contamination of the leachate. With increasing contents of readily degradable organic matter in the waste, the hydrolysis and acetic fermentation processes become faster and the equilibrium of complete methanogenic fermentation is further delayed. Although methanogenic bacteria in industrial processes with liquid media are observed to be highly sensitive, the same bacteria exposed to the solid medium of the waste are not so sensitive. High methane production rates are often observed under conditions that would seem to rule them out.

Because the content of organic matter in the waste is not unlimited, in the long term the above processes must cease. With neither CO_2 nor CH_4 being produced in significant amounts, air and thus oxygen can now diffuse into the waste fill, re-establishing aerobic conditions. This is not expected to occur in MSW landfills for centuries or millennia. If the ambient conditions change in this way, changes in leachate contamination and possibly increases in the concentrations of some constituents cannot be ruled out.

10.1.6.2. Water Balance

Because of the heterogeneous nature of the material and the granular structure, the behavior of water in a landfill is much more complex than in a natural soil body. This is especially true of landfills for municipal solid wastes. The coarse waste structure prevents uniform saturation of the fill, so that broad leachate flow paths can form. In addition, materials such as plastic film create barrier layers, paper stores water, and organic matter can absorb and distribute water as well as hinder its flow. Figure 140 illustrates the water balance. With regard to the total quantity of leachate, it is important how impermeable the waste surface is and how much precipitation can reach it. In many landfills, the quantity of leachate is reduced over a long time by water storage in the waste.

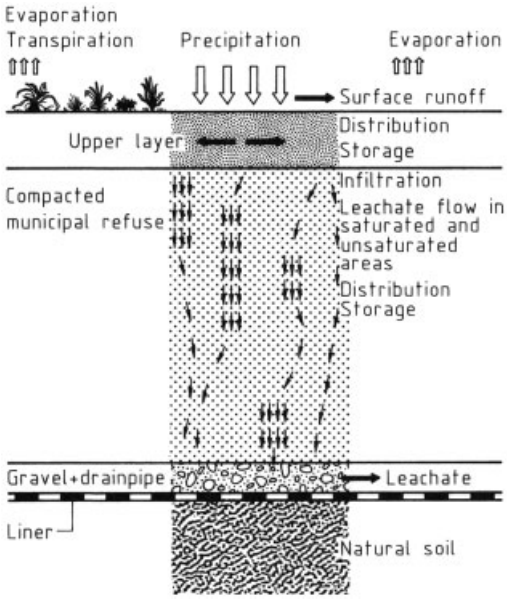

Figure 140. Schematic water balance of an MSW landfill

Leachate quantities observed at many landfills in the past (Fig. 141) were 10 – 25 % of precipitation [337], [339–341]. In very young landfills with high storage capacity, in contrast, leachate was less than 10 % of precipitation; in old landfills, after exhaustion of the storage capacity, it was as much as 40 % of precipitation. The figures given are yearly averages because the leachate discharge varies widely over the year. Figure 142 shows the leachate discharge of a landfill, with high values at the end of winter and spring and low discharges in late summer and autumn.

The storage capacity of waste has been significantly cut, above all through the separation of paper and the separate disposal of demolition waste and soil. Leachate rates of 30 –40 % of precipitation are often seen today as a result. At some landfills for MSW, the organic wastes are also separated for composting. The poor compactability of the remaining wastes together with their low storage capacity means that the leachate flow increases to 60 – 80 % of precipitation, with a particularly irregular discharge pattern [344].

At non-MSW landfills, the quantity of leachate is so strongly dependent on the type of waste and its compaction that rates from ca. 0 % to ca. 100 % of precipitation are possible. For example, industrial and hazardous waste landfills in the Federal Republic of Germany have recorded leachate discharges ranging from 10^{-5} to $1.86\,m^3\,m^{-2}\,a^{-1}$ [336].

10.1.6.3. Leachate Quality

Leachate quality is strongly dependent on the nature of the wastes and on the processes taking place in the fill (see Section 10.1.6.1). The measured values, of course, apply only to the initial period of a landfill, since there is not yet any experience with the disposal of wastes from our industrial society over hundreds or thousands of years. Because the substances placed in landfills and mobilized are not infinite, the emission stream must stop some time. It is difficult, however, to estimate when this will happen.

Table 31 lists the concentration ranges in leachates from several types of landfills in the EU [337], [340], [341], [344–350]. The landfill types in the table are not identical to those listed in Section 10.1.2, because the study concerned existing landfills (older than 3 – 5 years) and not new ones that come under the classfication scheme. As was mentioned in Section 10.1.6.3, MSW landfills often experience an acetic phase in which organic contaminant levels are high, which is later followed by a methanogenic phase in which organic contaminant levels are moderate. This progression can be seen in Figure 143, which presents data for an operated landfill.

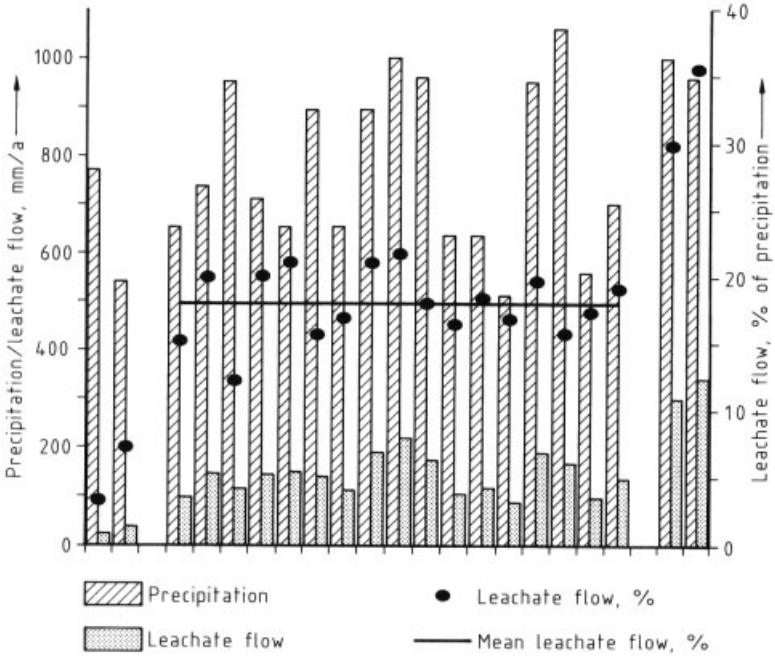

Figure 141. Precipitation and leachate quantities for various landfills

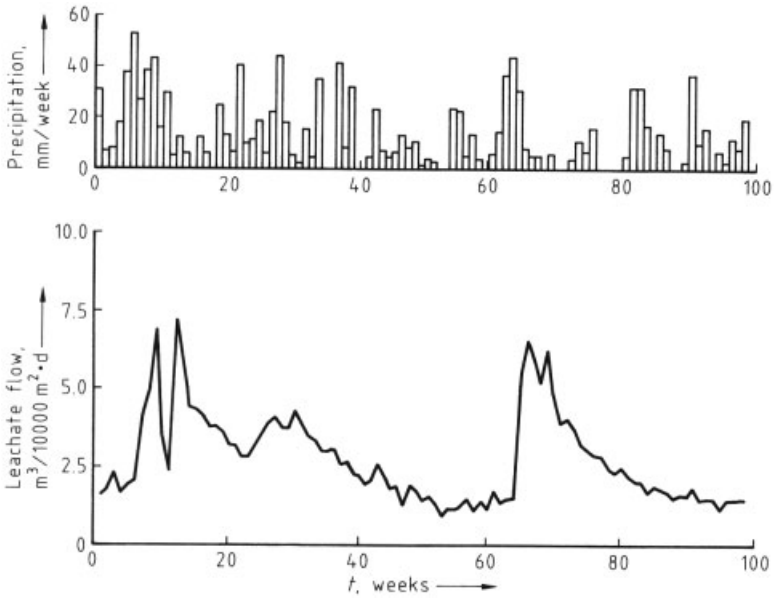

Figure 142. Precipitation and leachate flow for a company landfill

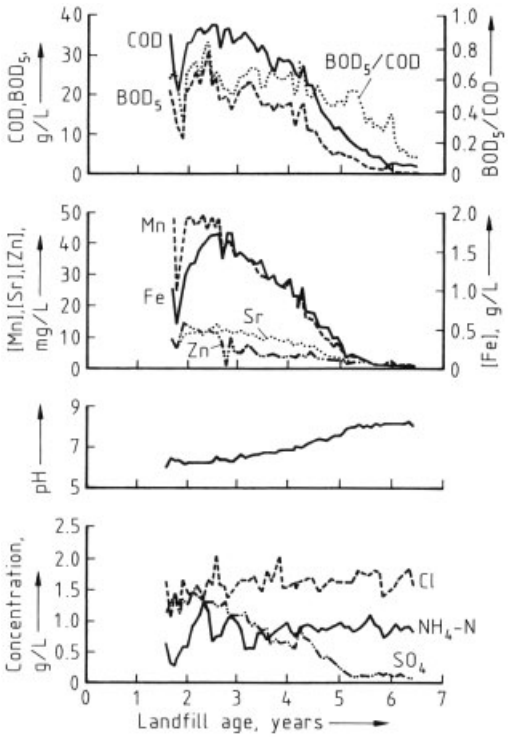

Figure 143. Concentration of various leachate components as a function of time

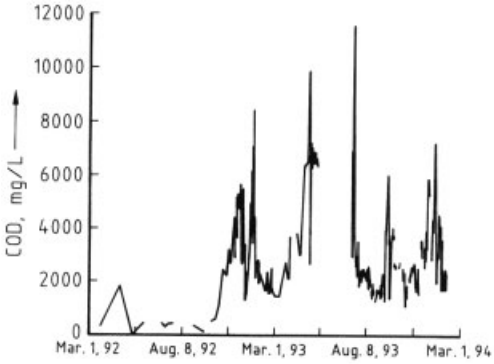

Figure 144. Variation in COD for a landfill with low water storage capacity of the waste

Because the organic contamination is due to volatile (lower) fatty acids, the pH falls and the concentrations of some inorganic substances such as calcium and iron rise markedly. Sulfate also undergoes changes due to biological processes, and these are energetically more favorable than methanogenic fermentation, so sulfate is largely reduced before full methane production is achieved. Care in the handling of problem substances (heavy metals, organic pollutants) has led to appreciable reductions in contamination of leachates with these substances. If smaller storage capacities result in very uneven discharges, large fluctuations in leachate contamination often accompany this effect, as shown in Figure 144 [344].

Operating techniques such as the placement of wastes in thin layers (depth 0.15 – 0.5 m) and recirculation of leachate have made it possible to reduce significantly the high organic contamination levels during acetic fermentation (Figure 145) [338], [341]. These measures prolong the initial aerobic phase and improve the water distribution.

Estimates indicate that water-polluting species such as nitrogen and organics (COD, BOD_5) present in leachate from many older landfills may necessitate treatment for 200 – 600 years [342], [343]. This time period may be much longer in the case of volatile organohalogen compounds (AOX), so that a reduction of the amount of these substances in wastes is urgently needed.

10.1.6.4. Leachate Treatment

Contaminants in leachates from landfills accepting MSW and some other, more highly contaminated wastes mean that these leachates cannot be released directly into surface waters. In Germany, the constituents relevant to treatment are organic matter (COD, BOD_5), ammonium, and AOX [351]. Often 400 – 1500 mg COD/L is not biodegradable and must be removed by physicochemical methods. The following combinations of processes are therefore employed:

1) Biological treatment to eliminate BOD_5 and nitrification and denitrification to remove nitrogen; flocculation if necessary to remove high molecular mass COD and AOX; adsorption on activated carbon to eliminate COD and AOX (with thermal regeneration of the activated carbon).
2) Biological treatment to eliminate BOD_5 and nitrification – denitrification to remove nitrogen; oxidation of COD and AOX with ozone (and UV radiation if appropriate); final biological treatment if necessary to remove biodegradable oxidation products.

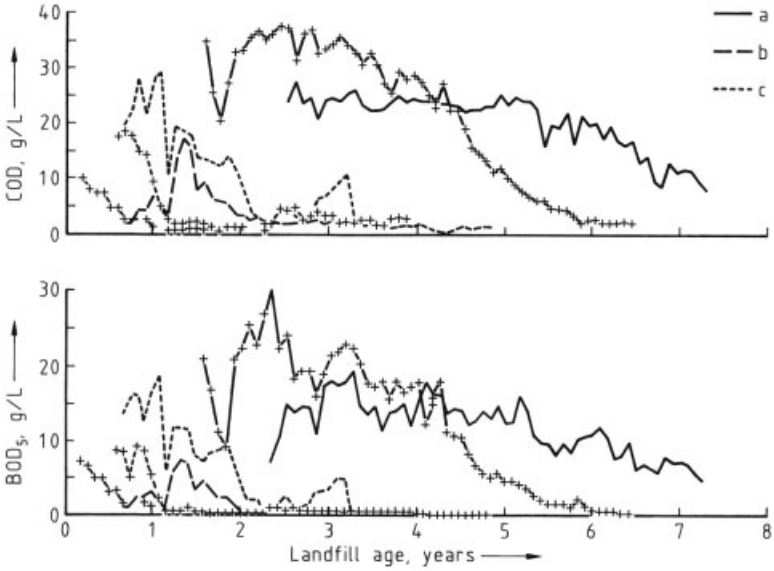

Figure 145. Typical COD and BOD_5 curves for operated landfills with different operational methods a) Deposition in 1.5 – 2 m layers; b) Deposition in thin layers; c) As (a) but with leachate recirculation

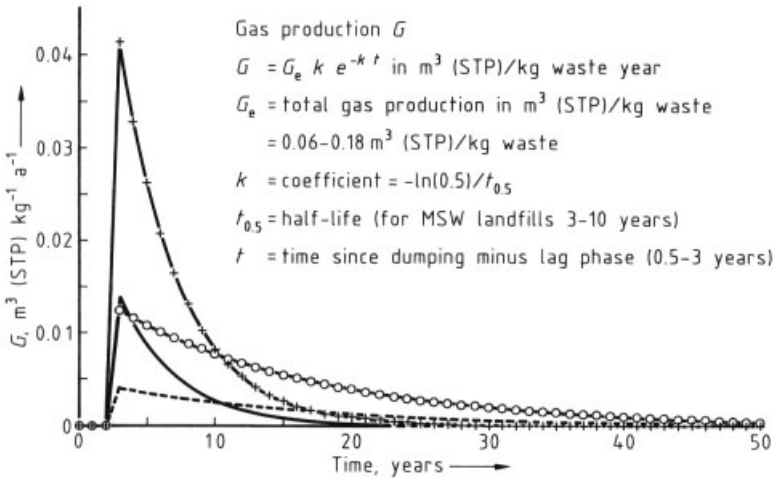

Figure 146. Variation of gas production with time for total gas quantity $G_e = 0.06 - 0.18\,m^3$ (STP)/kg and half-life $t_{0.5}$ of 3 and 10 years

3) Biological treatment if necessary to eliminate BOD_5 and nitrification – denitrification to remove nitrogen; stripping of ammonium if there is no biological pretreatment; one to three stages of reverse osmosis (depending on contaminant levels and required effluent values) to separate and concentrate nearly all substances contained in water; evaporation and drying of the concentrate and, if necessary, of ammonium sulfate liquor from stripping.

The third variant removes virtually all contaminants from the leachate but involves much technical effort and financial expenditure. The residues are readily soluble and therefore have to be stored underground to protect them from

weathering. In the first two processes the only residues are those from biological treatment and flocculation, if practiced; these present few disposal problems. While these processes offer lower costs than the chain of operations involving reverse osmosis, costs of leachate processing are still substantial.

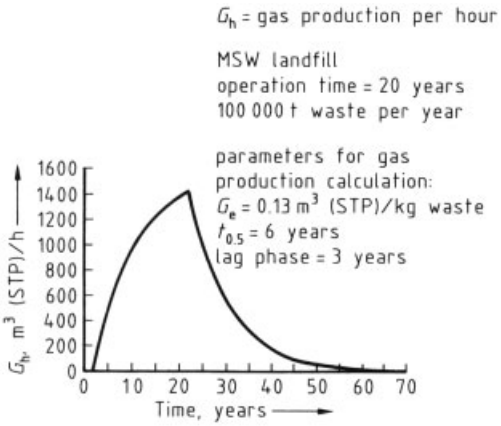

Figure 147. Variation of gas production with time for a landfill in which 100 000 t of waste were deposited over the course of 20 years

10.1.6.5. Gas Formation

The formation of gases in landfills can take place in several ways:

1) Evolution of volatile waste constituents (especially solvents)
2) Reactions between waste species resulting in gas production
3) Anaerobic conversion of organic substances by methanogenic fermentation (see Section 10.1.6.1)

Although it is known that the first two processes do take place in some industrial and hazardous waste landfills, little is known about them. Gas production by methanogenic fermentation is much more familiar because it occurs in all MSW landfills [337], [352–354]. Gas production begins spontaneously, some 0.5 – 3 years after deposition of the waste, and quickly reaches a maximum value from which it then declines slowly over a period of decades (Fig. 146). This gas production is fairly rigorously described by the decay function of Figure 147. The maximum quantity of gas produced is ca. 0.06 – 0.18 m^3/kg of waste (gas volume measured at STP). If the fraction of organic matter in the waste is lower, the total quantity of gas producible (G_e) decreases, and the half-life generally becomes longer. By way of example, Figure 147 shows the hourly gas output of a landfill in which 100 000 t of MSW per year was deposited over a 20-year span. The gas from MSW landfills consists of ca. 55 % CH_4 and ca. 45 % CO_2. Because the waste fill is not a closed system, air is often included in the measured gas flow; the above values then decrease. Less than 1 % of the gas consists of a wide range of trace substances, which (among other effects) render the gas odorous. Their concentrations in mg/m^3 are as follows [355]:

Ethane	0.8 – 48
Ethylene	0.7 – 31
Propane	1.4 – 13
Propene	0.04 – 10
Butane	0.3 – 23
Butene	1 – 21
Pentane	0 – 12
2-Methylpentane	0.02 – 1.5
3-Methylpentane	0.02 – 1.5
Hexane	3 – 18
Cyclohexane	0.03 – 11
2-Methylhexane	0.04 – 16
3-Methylhexane	0.04 – 13
Cyclohexene	2 – 6
Heptane	3 – 8
2-Methylheptane	0.05 – 2.5
3-Methylheptane	0.05 – 2.5
Octane	0.05 – 75
Nonane	0.05 – 400
Cumene	0 – 32
Bicyclo[3.2.1]octane-2,3-methyl, 4-methylene	15 – 350
Decane	0.2 – 137
Bicyclo[3.2.0]hexane-2,2-methyl, 5-methylethyl	12 – 153
Undecane	7 – 48
Dodecane	2 – 4
Tridecane	0.2 – 1
Benzene	0.03 – 7
Ethylbenzene	0.5 – 236
1,3,5-Methylbenzene	10 – 25
Toluene	0.2 – 615
m/p-Xylene	0 – 376
o-Xylene	0.2 – 7
Trichlorofluormethane	1 – 84
Dichlorofluormethane	4 – 119
Chlorotrifluormethane	0 – 10
Dichloromethane	0 – 6
Trichloromethane	0 – 2
Tetrachloromethane	0 – 0.6
Chloroethylene	0 – 264
Dichloroethylene	0 – 294
Trichloroethylene	0 – 182
Tetrachloroethylene	0.1 – 142
Chlorobenzene	0 – 0.2

Volatile chlorinated hydrocarbons are transported by the landfill gas.

The gas produced diffuses out of the landfill and can cause suffocation of plants, fires, and – in certain mixtures with air – explosions. The collection and disposal of this gas is therefore essential.

10.1.6.6. Landfill Gas Recovery

To avoid hazards in the operation of the landfill and for the surroundings, the gas produced must be extracted from the landfill. This is performed with collection pipes placed in the waste fill and surrounded by gravel. The pipes can be either vertical or horizontal. They must be carefully sealed in the outer parts of the landfill so that no air can enter the system. Mixing with air can lead to explosions. The gas is drawn from the collection pipes by suction (1 – 3 kPa) generated by a fan.

Because maximum gas production occurs as little as 0.5 – 3 years after placement of the waste, gas extraction during landfill operation is often necessary. This can be done more effectively and with less disruption if the gas recovery system is horizontal. It is still a matter of doubt, however, whether such systems will remain functional in the long term. Additional wells for vertical gas recovery systems are often constructed after the closure of a landfill.

The extracted gas is saturated with water because of the ambient conditions in the landfill. As soon as the gas cools after leaving the waste fill, water condenses from it. The water is frequently contaminated with high levels of volatile chlorinated hydrocarbons, so it must be separated and disposed of. Only when the contaminant levels in this condensate are the same as in the leachate can the two water streams be treated together.

The gas is destroyed in flare systems designed for high-temperature combustion; otherwise, the presence of chlorine in the compounds would pose the danger of dioxin and furan formation.

Landfill gas has an energy content of ca. 4 – 5 kWh/m^3 (not including conversion losses), and so using it for energy production may be feasible if the quantity is large. Some options are:

1) Direct combustion to generate heat
2) Combustion in motors or turbines to generate electricity
3) Removal of CO_2 and trace substances to obtain methane

The only forms of utilization that have been practiced on a significant scale are direct combustion with heat recovery and combustion in motors to generate electricity. Given typical landfill sizes, electrical capacities of 0.5 – 2 MW can be installed.

In the case of old landfills with low gas production rates, biological oxidation of methane in compost or soil filters is a possibility.

10.1.7. Landfill Operation

In the receiving portion of a landfill, arriving waste is commonly weighed and visually examined or inspected to determine its identity. The waste is then unloaded from the delivery vehicle in the operations section. A further visual identity check should be performed here. Next, the waste is incorporated in the fill one lift at a time. Inorganic wastes are often bulldozed into the fill. The height of the lifts often depends on the nature of the waste. Municipal solid wastes are compacted by heavy, steel-tired loaders (compactors). The wheels of these loaders have steel teeth welded on; the tips of the tooth generate a high peak pressure as the vehicle moves over the fill, compacting the MSW. In the past, such wastes were often placed in 1.5 – 2 m lifts and then compacted. It has been found that the pressure exerted by compactors is not enough to optimally compact this high a lift. Today, lift heights at many landfills are 0.15 – 0.5 m.

Landfill operators are responsible for protecting the environment from dust and odors. High-growing vegetation is often planted around the landfill boundaries as a visual shield.

10.2. Underground Deposition

10.2.1. Introduction

Progress in underground construction in recent decades means that the building of underground cavities today is economical, safe, and competitive with aboveground structures in many

Table 31. Concentration ranges in leachates from various types of landfill

Parameter	Demolition waste landfills	Monolandfills (for incinerator residues)	Industrial waste landfills	MSW landfills	MSW landfills	Co-disposal landfills	Hazardous waste landfills
				Acetic phase	Methanogenic phase		
pH	7 – 8.1	7.2 – 10.5	6.9 – 9.7	6.1	7 – 9.8	7.4	5.9 – 11.6
Na, mg/L	104 – 270	1600 – 7300	110 – 1300		1350	1513	
K, mg/L		600 – 4300	90 – 3600		1100	603	
Ca, mg/L	200 – 335	32 – 1000		1200	60	156	
Mg, mg/L	30 – 47	0.09 – 66		470	180	176	
total N, mgN/L		2.4 – 100	9.1 – 480		1250		
NH_3, mgN/L	13 – 28	1 – 254	0.08 – 565		750	665	<5 – 6036
NO_3, mgN/L		0.04 – 1594	160 – 780		3		<0.1 – 153
NO_2, mgN/L					0.5		0.02 – 131
total P, mgP/L		0.08 – 1.4	0.03 – 1.41		6		0.03 – 52
SO_4, mg/L	<5 – 450	55 – 7200	286 – 3630	500	80	113	18-14 968
Cl, mg/L	100 – 600	160 – 87 000	320 – 12 300		2100 (1000 – 2000 ∗)	2131	36 – 126 300
BOD_5, mg/L		<1 – 1700	0 – 1300	13 000	180	674	<5 – 15 000
COD, mg/L	130 – 236	75 – 3530	17 – 2240	22 000 (5000 – 12 000 ∗)	3000 (800 – 1500 ∗)	2366	50-35 000
TOC, mg/L	40 – 93	6 – 44		8500	1050	661	
AOX, µg/L			200 – 6700		2000 (500 – 1500 ∗)	340	40 – 292 000
EOX, µg/L	10 – 20						
phenols, mg/L	<0.02		0.04 – 3.4		<1	0.45	<0.01 – 350
Fe, mg/L	3.5 – 260	<0.01 – 18.5	0.21 – 142	780	15	16	0.21 – 2700
Zn, mg/L	0.1 – 06.	<0.01 – 19	<0.01 – 7.2	5	0.6	0.82	<0.02 – 27.2
Pb, µg/L	0.2 – 6	<0.5 – 900	5-2600		90	110	4.3 – 1000
Cd, µg/L		<0.1 – 1500	0.54 – 160		6	max. 30	<0.2 – 2000
As, µg/L	9 – 37	5 – 25			160		<2 – 240
Cr, µg/L	3 – 8	<1 – 100	10 – 120		300		<9 – 300 000
Cu, µg/L	1 – 11	<0.5 – 300	30 – 300		80		1.3 – 8000
Ni, µg/L		<2 – 600	50 – 3200		200		14.2 – 30 000
Hg, µg/L		<0.05 – 106	<0.2 – 1.8		10		0.17 – 50

∗ Newer landfills with separation of refuse.

respects. Furthermore, the use of underground cavities offers certain advantages over surface facilities, particularly with respect to maintenance, operation, and safety. Underground storage for petroleum products and gaseous hydrocarbons has proved incomparably cheaper than freestanding storage.

Underground waste deposition has become increasingly important in the waste management field. The reasons lie partly in the decline of mining. Strong international competition in the extraction of raw materials has led to a search for new, less costly sources. In Germany, some mining companies have begun offering to dispose of wastes in their deep or surface mines in parallel with their mineral extraction programs. In addition, stiffening competition on the international raw-materials market has forced the closure of many German mines. As a result, much underground capacity suitable for waste storage and disposal has become available. In a few instances, underground cavities have been constructed specially for waste disposal; this is particularly so in the field of ultimate disposal of radioactive wastes. Underground waste disposal and recycling has grown into a new economic sector in recent years, and development in the field is rapid, driven by changes in the waste and residue market, the oversupply of unused surface and deep mines, and also by changing conditions of ownership. There is a huge stock of exploitable tailings and spoil piles as well. The market can be expected to change because the legal situation is not clear at the moment.

The main aspects of underground deposition in Germany are covered by TA Abfall and var-

ious waste management and mining laws. The government ministries with jurisdiction in this area, as well as the regulatory authorities dealing with mining and waste, are now concentrating their efforts on closing the gaps that still exist in the laws and implementing regulations. The creation of the European single market, however, suggests that more changes are in store for this sector.

10.2.2. Types of Mines and Their Relevance to Waste Deposition and Recycling

Essentially there are three ways of depositing wastes in mines:

1) The use of existing cavities created by mining raw materials
2) Construction of new mines specifically for waste disposal in suitable geological formations
3) Injection of liquid residues and wastes into natural cavities, such as pores and joints in subterranean formations (e.g., exhausted gas and oil fields)

Three forms of mining are practiced: surface (open-pit, open-cast, strip) mining; underground (deep) mining; and borehole mining.

In surface mining, deposits lying under shallow cover near the surface are worked where the overburden can be economically removed with existing technology. In Germany, lignite, rock, and earths are mined in this way. As a rule, surface mines are carried to depths of 100 – 200 m. Deeper pits, such as the lignite mines in the Aachen district, extend to depths of 400 – 500 m. In the past, surface mines were often used for waste disposal. Germany has countless worked-out pits and quarries where clay, sand, gravel, and stone were produced, as well as closed lignite mines, and in the past these were used – sometimes casually – as dumps. Today, because of the lack of foundation seals and other defects in water-retention and drainage practices, some of these dumps have become major hazardous waste sites that will require major efforts to clean up.

Most surface-mine sites are used for the disposal of mining wastes and the replacement of overburden. Certain residues, such as ash, slag, and gypsum derived from flue-gas desulfurization at lignite-fired plants, can be included with this material with no detrimental impact on groundwater. Lignite ash has a buffering action on the sometimes very acidic deep groundwater and may actually have a beneficial effect on the groundwater situation. Surface mines where clay was extracted (e.g., clay pits and brickyards) used to be employed as hazardous waste landfills. Because the dense clay formations were mostly removed in the mining operation, these cavities often lack an adequate foundation seal, so that isolation from deeper geological formations can no longer be guaranteed. Gravel and sand deposits commonly occur in highly permeable groundwater horizons near the surface and are unsuitable for the deposition of any kind of wastes except completely inert materials. Stone quarries are generally unsuitable for waste disposal too, because of the jointed character of the rock. If they are used for this purpose, technical modifications such as liners at the base and sides must be employed to provide the necessary impermeable barrier between the waste and the geological surroundings.

In subsurface mining, raw materials are won by driving cavities into the geological formation, and the overburden is left in place. Deep mines require accesses for transportation, ventilation, and drainage. These accesses take the form of galleries or shafts, which may be inclined or vertical, depending on the form of the deposit. In Germany, coal, potassium compounds, rock salt, and ores are extracted by deep mining.

Borehole mining is a special form of mining, mainly used for petroleum and natural gas production. This technique is best suited to liquid and gaseous raw materials, which may lie at great depths. The variant of solution mining is used for salt deposits. The cavity left behind by this type of mining is often utilized for gas and petroleum storage.

The use of mine cavities for the deposition of residues or wastes is governed by one of three laws, depending on the nature of the material: the Mining Act (Bundesberggesetz), the Waste Act (Abfallgesetz), or (for radioactive wastes) the Atomic Act (Atomgesetz).

When wastes generated in the prospecting, production, beneficiation, and refining of raw materials and subject to inspection by the mining authorities are to be replaced in mine cavi-

ties, they fall under the German Federal Mining Act (Bundesberggesetz) [356] and must be covered by an approved operating plan. Mineral residues not derived from mining can also be used in stowing.

If wastes and hazardous wastes are to be disposed of, all facilities, mines, caverns, and surface workings are subject to mandatory approval under applicable eminent domain provisions of waste-management law [357]. The deposition of radioactive wastes in mines, caverns, or boreholes is regulated under the Atomgesetz [358]. Facilities used for this purpose are again subject to approval; they simultaneously fall under supervision by the mining authorities and require the submission of a mining operations plan for supplemental approval. A regulation called Technical Directive on the Disposal of Hazardous Waste (TA Sonderabfall) [359] has been issued to cover the deposition of wastes in mines; it sets forth requirements on the quality of the underground facilities as well as their surroundings, the disposal method, the waste spectrum, and other points. An interstate working group on mining in Germany also published an engineering code relating to the disposal and reclamation of mineral raw materials and wastes used as underground stowage [360].

10.2.3. Safety Requirements for Operation and Closure of an Underground Deposition Site or a Mine Subject to Stowage

Mine cavities used for the disposal of wastes or the storage of residues are classified in five categories; the principal parameters are the form of the deposit, the groundwater situation, and the mining method. TA Abfall deals with the Type 1 underground deposition site, the salt mine. Here the barrier between waste and surroundings is the salt formation. If the salt is not very thick, additional layers in the overburden are required for sealing against groundwater. Such a working is sealed by lining the shaft where it passes through the groundwater horizon. The waste deposition rooms are sealed by closure systems such as embankment dams and also by the shaft closure required after the operational phase. As far as the technical parameters of the cavities during operation are concerned, stability must be ensured. The wastes themselves must not react with one another or with the host rock, so that no gas evolution or dissolution of the salt can occur. Underground deposition sites of this type are already in service or planned in the German states of Baden-Württemberg, Hesse, North Rhine-Westphalia, Lower Saxony, Saxony-Anhalt, and Thuringia.

A variant of Type 1, salt mine, is underground deposition site Type 2, the salt cavern. This cavity is likewise isolated from its surroundings by impermeable salt formations that guarantee safe containment of the wastes. In contrast to Type 1, this structure is created by solution mining. Access to the surface is through a borehole. Accordingly, the wastes must be placed in the salt cavern through a borehole; they can be delivered by free fall, pumped, or pneumatically conveyed. Unlike the wastes for Type 1 underground deposition, which can be either packaged wastes or unpackaged stowage materials in the form of slurries or the like, wastes for Type 2 deposition are not in containers. Because of intensive salt mining in Germany, large volumes are available for waste deposition and there are opportunities for the stowage of residues; this is expected to grow into a self-sustaining disposal market in coming years. Like Type 1, Type 2 is also covered in TA Abfall.

Underground deposition facilities of Types 3, 4, and 5 are not dealt with in TA Abfall [361].

Type 3 deposition involves mines in non-aquifer formations. Such a mine is driven in a geological formation that may be permeable but does not have any groundwater flow either because of impermeable roof rocks or as a result of special tectonic conditions. The underground cavity is sealed off from the aquifer by a tight layer; access is via a shaft, which must be sealed with a shaft plug after the end of the operational phase. The rooms in which wastes are deposited can also be closed off by embankments. As in the salt mine, a stability analysis must be performed during the operational phase, and on-site personnel must be in charge of ventilation and must inspect the residues and wastes from an occupational-health standpoint. Wastes can be deposited in containers, pumped in through pipes, or conveyed by pneumatic systems or belts. As in the case of the salt formations, wastes must not evolve gas. Mines of this type, chiefly in ore-producing regions, are being prepared for underground deposition service.

Type 4 underground deposition facilities are mines in aquifers. During the operational phase, the mine rooms are kept dry by drainage systems. After these systems are shut off, the wastes and residues are located in the groundwater. Mines of this type are used for stowage chiefly in coal-mining regions. Care should be taken that the elution behavior of the materials – which arrive at the disposal site chiefly as concentrated and dilute slurries and are stabilized on-site by the addition of pozzolanic cements – is such as to prevent appreciable releases of contaminants.

The Type 5 category of underground deposition sites is made up of mines above the aquifer, which generally are dry. Under some circumstances, however, precipitation water must be expected to penetrate into the roof rocks or the mine itself, so that permanent isolation of the wastes from the groundwater cannot be guaranteed. In mines of this type, wastes are deposited by on-site personnel; as in the Type 4 case, care should be taken that the wastes, usually consolidated with pozzolanic cements, are immobilized so that contaminants will not be eluted into the groundwater.

A special form of waste disposal is injection into natural pore or joint volume. The petroleum and natural gas industry commonly uses exhausted oil and gas fields for the storage of formation water produced along with the oil and gas. At some German sites, however, waters not originating in field operation – mainly saline waters from salt refining – are injected into deeper-lying joints and pores. As a rule, the geological formations are isolated from exploitable groundwater by impermeable roof rocks.

When constructing underground deposition sites, care should be taken that contamination of the atmosphere and hydrosphere and detrimental changes in the surroundings of the site are avoided. This essentially means testing the geology and hydrogeology of the site and of the underground space considered as a barrier system. Engineered safety systems must be used to secure accesses with suitable closure systems. The regulatory approval procedure calls for a safety analysis based on safety testing of the entire system, together with accident analyses. In addition, the wastes to be disposed of must be rendered compatible in the sense of disposability and possible chemical reactions with the barrier systems and the enclosing rocks.

10.2.4. Wastes and Residues for Deposition in Mined Cavities

Wastes deposited in underground cavities must be brought into a form suitable from the standpoint of geotechnology, occupational safety and health, and transportation. One way of achieving this is to package the wastes in containers; another is to subject them to physicochemical or, more often, thermal treatment.

The spectrum of wastes suitable for underground deposition essentially takes in two major waste categories:

1) Wastes of mineral origin and those stemming from refining processes; examples are metallurgical slags, ashes, slags and dusts, mineral sludges, and nonferrous metal wastes
2) Wastes from conversion and synthesis processes; examples are salts, oxides and hydroxides, pesticide wastes, pharmaceutical manufacturing wastes, petroleum-based greases and waxes, coating materials, and catalysts

Not suitable for underground disposal in accessible cavities are wastes that have a penetrating odor, give off-gas, or form fine dusts that can be carried by ventilation air and present a hazard to the mine personnel. Wastes that ignite spontaneously or sustain combustion, explosive wastes, and wastes that form toxic or explosive gas – air mixtures under the conditions prevailing in the mine rooms are also excluded. This means that wastes whose organic character makes them liable to decompose and endanger the health of the mine personnel are essentially not eligible for this kind of disposal. The spectrum of stowable materials and residues includes wastes from coal-fired plants, slags and ashes from thermal treatment of wastes, residues and wastes from foundries, construction wastes, wastes from reaction and synthesis processes, and abrasive residues. Before wastes can be deposited or residues recycled, a substantiated description of what is planned must be submitted by way of an assessment. This declaration must take the form of disposal and recycling certificates in the case of residues and wastes subject to regulatory supervision. For other residues and wastes, the requirement of a disposal certificate

holds as appropriate. The monitoring of reclamation practices includes inspection of manifests on delivery, a comparison of the delivery manifest with the recycling certificate, a determination of the quantity, and an identity check. Pursuant to mining law, the authority is responsible for monitoring the disposal procedures in cooperation with other regulatory bodies.

10.2.5. Examples of Spaces Used for Waste Deposition and Spaces Packed with Residues Not Generated in Mining

Hazardous wastes have been placed in the Herfa-Neurode underground depository of Kali and Salz since 1972. The plant has been designated an underground waste deposition site and licensed under waste management law. This facility is part of a mining territory of the Wintershall natural potassium salt mine, which is still producing the salt. As parts of the pillar-and-stall working are closed, rooms are bulkheaded off and used for the deposition of hazardous wastes. Since mining continues in the complex, wastes must be packaged in containers. The prerequisite for disposal use is the possibility of isolating the closed and decommissioned mining territory from the producing territory. Cavities made by mineral mining must be stable. The mine in which the wastes are deposited must be dry and absolutely water-free. The geological conditions comprise a Zechstein salt formation ca. 300 m thick with horizontal bedding overlain by a 300 – 600 m thick roof. Further impermeable horizons in the form of clay beds some 100 m thick are found in the overlying rock; these seal the mine off from higher formations and groundwater.

Before the deposition of wastes, the closed cavities are upgraded by safety measures such as roof anchoring, removal of loose material, and treatment of the floor for vehicle access. The underground disposal site was approved for this service by the Hessian mining authorities in accordance with waste management law. The construction and operation are governed by regulations implementing the Abfallgesetz as well as provisions of mining law. The Hessian Bergamt (mining inspectorate) is the supervisory agency for the underground disposal site; this body includes other authorities and institutions in its approval procedure, such as Hessische Industriemüll GmbH, the Hessian Bundesanstalt für Umwelt, and the Hessian environmental ministry.

On delivery, wastes are classified by substance class and placed in the appropriate areas. In all, 16 substance classes are stored in separate rooms; the major classes are water-hardness salt residues, electrolytic cell muds, chemical residues, drying agents, and residues from flue-gas cleaning systems at hazardous waste and domestic waste incineration plants.

The wastes are generally packaged. It is also required that the wastes be neither explosive nor spontaneously combustible. They must be solid, must not contain any free liquid, and must not evolve gases. Radioactive wastes are excluded. Wastes are inspected and sampled before deposition.

The second underground waste disposal site in Germany is the Heilbronn salt mine of Südwestdeutsche Salz, which received approval in 1986 for the disposal of residues from flue-gas cleaning units at domestic and hazardous waste incineration plants. Up to mid-1994, a total of 200 000 t of flue-gas cleaning residues had been deposited in the site. The mine was built to extract salt from a Middle Muschelkalk salt formation ca. 40 m thick; it lies at a depth of ca. 170 – 210 m. Wastes are packaged in either sheet-metal containers or big bags and laid one on top of another in several lifts until the rooms are completely filled. After each room was filled with ca. 15 000 t of flue-gas residues, access was closed off with a wall and in this way the room was isolated from the ventilation system.

Germany's third underground disposal site, the Zielitz mine of Kali und Salz, is located in Saxony-Anhalt State. As at Herfa-Neurode, potassium salt is mined in board-and-stall fashion in a horizontally bedded formation. Closed and mined-out rooms are used for waste disposal. After individual rooms are filled, they are sealed off from the ventilation system with walls. The facility, licensed under waste management law, will begin disposal operations in 1995.

Two other sites are preparing for the regulatory approval process for hazardous waste disposal. These are the Riedel mine of Kali und Salz in Lower Saxony and the Borth mine of Solvay in North Rhine-Westphalia. The Borth mine is a pillar-and-stall working in horizontally bedded

Zechstein salt formations; the Riedel mine is in a salt stock where both mineral salt and native potassium salts are extracted by shrinkage stoping.

In addition to waste disposal in mines approved for this purpose, residues are increasingly being used in stowage. When deposits in mines have been driven through and mined out, mining law requires that they be rendered safe. In many such locations, long-term stability and freedom from damage at the surface cannot be guaranteed, so the mining authorities have decreed that they must be packed. To lower the costs of the mine operators, the mining authorities are increasingly allowing materials not originating in the mining operations to be used in stowage; the mine operator commonly realizes some income by accepting such "foreign" material.

One of the oldest packed mines where foreign fly ash from coal-fired power plants was used in stowing is the Wohlverwahrt-Nammen mind of Barbara Rohstoffbetriebe in North Rhine-Westphalia. This iron ore mine, worked by the pillar-and-stall method, continues to produce ore. For many years, power-plant fly ash, which has pozzolanic qualities, has been incorporated in the stowage to ensure the long-term stability of mined-out rooms until the mine is finally closed.

The Bernburg mine of Mitteldeutsche Salzwerke GmbH in Saxony-Anhalt state, also a pillar-and-stall salt mine, has accepted foreign stowage material for years in order to stabilize its salt pillars in some areas.

A pilot stowing program is undergoing tests in the Kochendorf salt mine of Südwestdeutsche Salzwerke. A similar pilot project is in operation at the Bleicherode mine of the Gesellschaft zur Verwertung und Verwahrung von Bergwerkseigentum (GVV) in Thuringia.

In the coal-mining sector, a pilot study aimed at reducing surface damage is being carried out at the Zollverein mine, where the caving method was employed. Foreign residues are being successfully used in stowing. Stowing of caved rooms with foreign material is also being practiced at the Haus Aden/Monopol mine at Bergkamen, Hugo Consolidation at Gelsenkirchen, and Walsum at Duisburg.

The closing of other mines is expected to generate a new market in stowage technology; it is anticipated that the method will gain approval under both mining and waste management law.

Now in use for the disposal of radioactive wastes is the Morsleben salt mine, which was licensed under the nuclear regulations of the former German Democratic Republic. This mine continues to be used for the deposition and disposal of radioactive wastes under a transitional regulation adopted at the time of German reunification, but it must be re-approved in the next few years.

In the context of a regulatory approval procedure, the Konrad ore mine at Salzgitter is currently being tested as an ultimate disposal site for low-level to medium-level radioactive wastes.

The Gorleben exploratory mine is now being driven in Lower Saxony state. Already approved under mining law, the working is intended to acquire crucial data needed for a proof of concept in the nuclear regulatory approval process for an ultimate disposal facility for highlevel and medium-level radioactive wastes.

10.3. Sealing of Existing Sites

Cutoff Walls. Landfills and abandoned sites often pose dangers to groundwater if appropriate construction techniques (foundation liner, slope seals, etc.) were not employed. To protect the groundwater against possible contamination from these sites, many rehabilitation projects involve constructing cutoff walls around the source of contamination. A cutoff wall has its toe keyed into impermeable soil layers so that it forms a trough-like containment (Fig. 148).

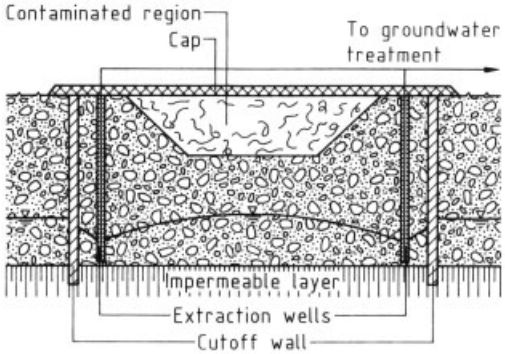

Figure 148. Containment with cutoff walls (thin wall)

Cutoff walls are classified by construction method:

1) Vibrating-beam cutoff walls (thin walls)
2) Slurry walls
3) Soil-mixed walls

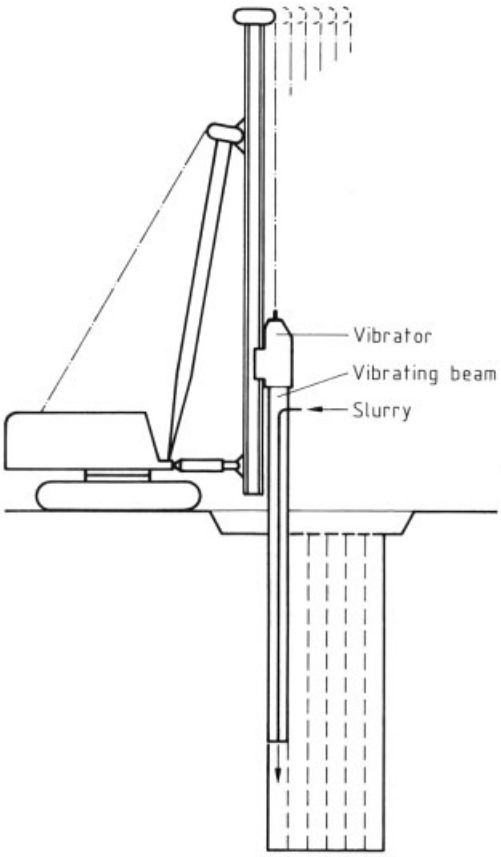

Figure 149. Construction of a vibrating-beam cutoff wall

A vibrating-beam cutoff wall is made with vibrating piles; in the other types, an excavator or cutter is used. Vibrating-beam cutoff walls are typically ca. 8 cm thick; maximum depths are around 20 m, and up to 30 m in favorable soil conditions. In building a vibrating-beam wall, the pile is driven to the design depth including ca. 1 m for keying into the impermeable stratum (Fig. 149). As the vibrating pile is withdrawn, the slurry issues at high pressure from the nozzle located at its toe, so that the cavity left by the pile is gradually filled with the cutoff wall slurry. The next cavity is made to overlap the previous one in order to prevent defects in the wall.

1 m^3 of slurry used in a slurry wall usually consists of

25 – 40 kg	bentonite (powdered)
150 – 250 kg	blast-furnace slag cement (HOZ 35 L)
900 – 940 kg	water

In the excavated trench, the slurry first stabilizes the trench itself; afterwards, as setting proceeds, it forms a seamless, impermeable wall. Prepared mixtures with activated bentonite are also available for making cutoff walls. These can be used immediately after dispersing with water and are mainly used in small restoration projects.

Normally, the specifications for restoration work on existing sites call for a permeability $k \leq 10^{-9}$ m/s. If this value is achieved or bettered, the cutoff wall is said to be "structurally tight." An excavated trench wall is also built with a specified amount of overlap between the vertical "lamellae," which depends on the design depth of the trench. The depth of keying in the impermeable bed, generally a clay bed, can be monitored by means of the excavator depth. The approximate topography of the stratum can be determined fairly precisely by borehole tests performed before the construction of the wall proper.

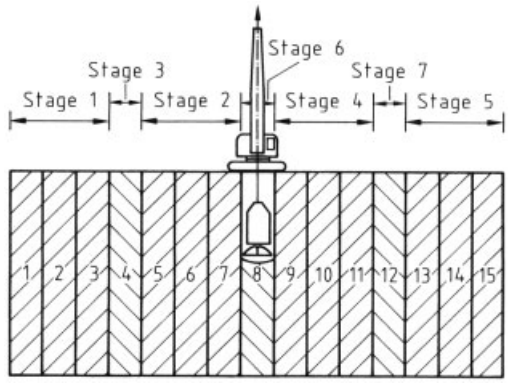

Figure 150. Construction of a cutoff wall by excavation

The following details should be kept in mind during construction:

1) The excavator must be guided vertically
2) The cutoff wall must be keyed into the impermeable stratum
3) The composition of the bentonite-cement slurry must be monitored
4) The attained permeability must be measured

5) The overlap must be checked

Trenches up to 35 m deep can be constructed with the excavator (Fig. 150). If subsoil conditions are favorable, the depth can be extended to 40 m. When even greater depths are desired or ground conditions are poor, as for example when lime benches are present, a rotary cutter can be employed (Fig. 151).

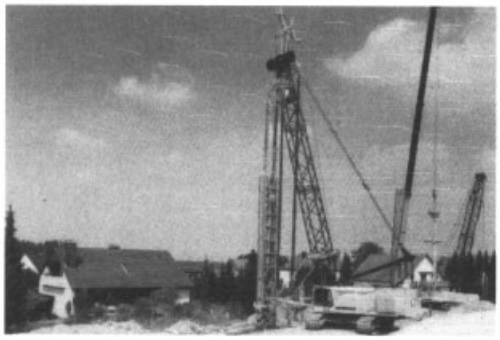

Figure 151. Construction of a cutoff wall with a cutter

The composition of the leachate may lead to impairment or destruction of the cutoff wall slurry. In such cases, cement-free mixtures (e.g., silicate-based) can be used. The initial excavation is carried out in the same way as for a slurry wall; the supporting liquid is then replaced by the wall mixture ("contractor" method). This cutoff mixture is very high in solids (around 90 %), with only the remaining 10 % consisting of liquid. The solid phase of such a cement-free wall is made up of sand and gravel as well as a nonswelling clay to perform the sealing function. The gel-forming liquid phase consists of water, waterglass, and two Dynagrout components. In order to improve safety during construction and operation, cutoff walls can also be built on the "chamber" system (Fig. 152). Parallel cutoff walls (vibrating-beam or slurry or a combination) are constructed around the contaminated region at an interval (ca. 8 – 10 m) dictated by the equipment. Every 50 m, say, or as far apart as 80 m, the cutoff walls are connected by transverse walls made from the same material, so that separate cutoff chambers are obtained. Each chamber has a monitoring well where continuous measurements of water level and water quality can be performed. In this way, the customer is enabled to test the functionality of the containment during construction and also to monitor the wall system during subsequent operation and make certain the cutoff wall system meets requirements. In case of a failure, the defect in the wall can be localized fairly quickly, and the failed chamber can then be repaired.

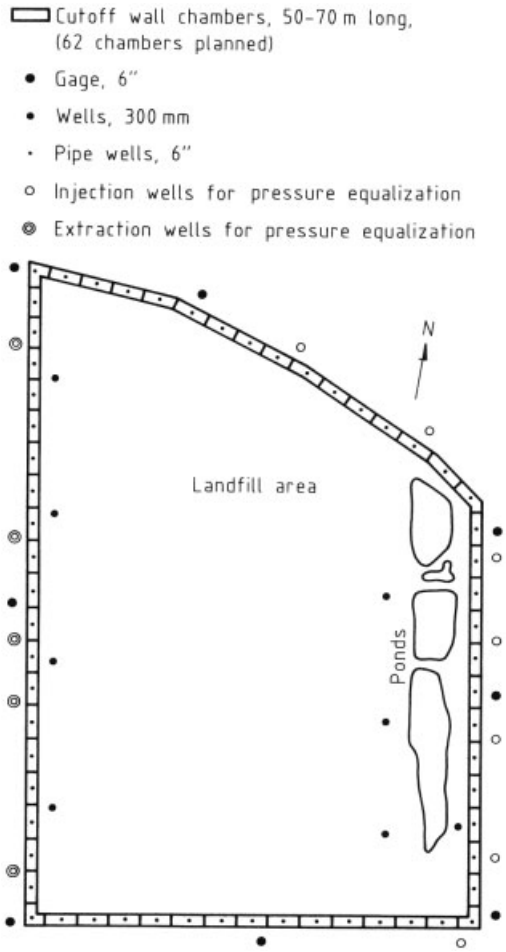

Figure 152. Chamber system of cutoff-wall containment

The safety standard for containment systems can be further enhanced by hydraulic systems. If the groundwater level inside the cutoff wall is made lower than the level outside the wall, then the operator can be sure that groundwater can move from outside to inside but under no circumstances can contaminated leachate escape from the contaminated region into the groundwater (Fig. 153).

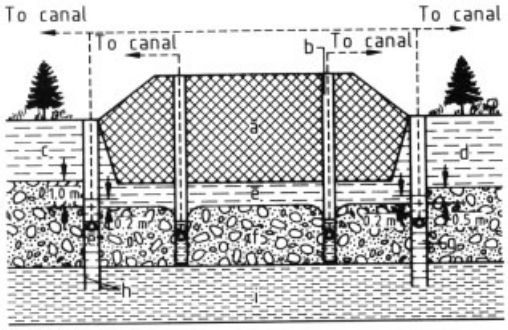

Figure 153. Hydraulic aspects of containment of an existing site (Rautenweg landfill, Vienna)
a) Waste; b) Well, 300 mm diameter; c) Groundwater, upstream; d) Groundwater, downstream; e) Groundwater; f) Sand/gravel; g) Cutoff chamber; h) Double cutoff wall; i) Aquiclude

When safety requirements are more stringent – that is, when the leachate discharge is to be eliminated or when the contaminants present in the leachate are aggressive – composite cutoff walls are built. High-density polyethylene (HDPE) liner material is installed in the slurry while it is still fresh. The liner (standard width 5.00 m) is delivered to the construction site and cut to length, and HDPE connectors are attached to the long edges. The liner material can be stretched over steel frames or unrolled from long rolls and inserted into the excavated trench while the slurry in the trench is still liquid. HDPE is used because it has proved to withstand chemical attack better than other materials. The liner must meet strict physical or physicochemical standards.

There are other engineered alternatives for the containment of existing waste sites, depending on the objective, the geological situation, the composition of the leachate, and other factors. The overlapping cast-in-place concrete pile wall is an example.

Capping Systems. Lining systems known as final covers are constructed to seal hazardous waste sites and also after the closure of landfills. They prevent precipitation percolating into the waste fill and leaching out contaminants. The design varies, depending on the nature and severity of contamination present and, in particular, the solubility of the contaminants in water.

In the case of relatively unobjectionable fills such as construction rubble, slightly contaminated soils, and slightly contaminated power-plant residues, an inorganic material (e.g., clay) provides adequate capping; see Figure 154. If the wastes or residues contain no organic material, the installation of gas drainage can be omitted. The seal system consists of two layers; for reasons having to do with construction technique, this system should have a maximum thickness of 25 cm. In order to produce a ground that is as homogeneous as possible and has the best loadbearing capacity, an equalizer layer is often placed over the contaminated region. The seal proper is overlain by a drainage layer, which should allow prompt removal of water percolating through the recultivation layer. The use of a geotextile between the drainage layer and the seal layer is recommended, and likewise between the drainage layer and the recultivation layer, to ensure the ability of the system to remove water.

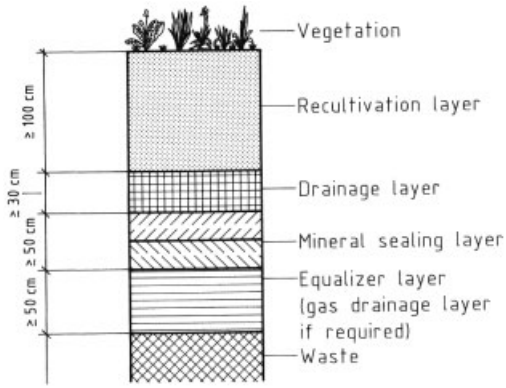

Figure 154. Capping system (landfill class I under the TA Siedlungsabfall in Germany)

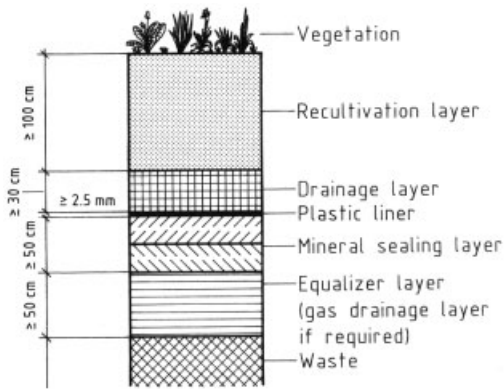

Figure 155. Composite capping system (landfill class II under the TA Siedlungsabfall in Germany)

If it is necessary to prevent rainwater from entering the fill, composite capping systems are employed (Fig. 155). Such a capping is made up of a mineral seal layer, supplemented by HDPE liners laid directly on the surface of the seal layer. The strips of plastic liner are usually welded by the heated wedge method. The welding of the double seams, with an inspection duct, is done by machine.

The system will be optimally effective if there is an intimate bond between the plastic liner and the mineral seal layer. If the surface is not highly inclined, this is fairly easy to achieve, but in steeper slope regions the construction may present some problems. It is recommended that such cappings be placed only during the warm summer months.

Among the important points for the long-term functioning of a capping system is the maintenance of a relatively high gradient at the surface (e.g., 5 %). This slope should still be present after any settlement has ceased, so that the rapid removal of precipitation can be guaranteed on a permanent basis. The minimum thickness of the recultivation layer is 1 m, though in special cases it may be much greater, for example if a visually appropriate match to the surrounding landscape is desired or if deep-rooted plants are to be grown.

11. References

1. M. Lehmann: *Abfalltechnik,* Vorlesungsmanuskript, Dübendorf 1992, 256 pp.
2. H. van Wickeren: *Mechanisch-biologische Behandlungsanlagen für Restmüll – eine Übersicht,* Der Städtetag, pp. 104 – 105.
3. Statistisches Bundesamt (ed.): *Statistisches Jahrbuch,* Wiesbaden (published annually).
4. K. Cord-Landwehr: *Einführung in die Abfallwirtschaft,* B. G. Teubner, Stuttgart 1994.
5. B. Bilitewski, G. Härdtle, K. Marek: *Abfallwirtschaft,* 2nd ed., Springer Verlag, Berlin 1994.
6. J. Wuttke: "Mengenstruktur beim Sonderabfallaufkommen," in: *Reglementierte Märkte blockieren den Wettbewerb,* Dokumentation zum Entsorga-Congress, Köln 1994.
7. K. J. Thomé-Kozmiensky: *Kreislaufwirtschaft,* EF-Verlag für Energie- und Umwelttechnik GmbH, Berlin 1994, 1170 pp.
8. Der Rat von Sachverständigen für Umweltfragen: *Umweltgutachten 1994 – Für eine dauerhaft-umweltgerechte Entwicklung,* Metzler-Poeschel, Stuttgart 1994, 380 pp.

General References

9. H. Birn, G. Jung: *Abfallbeseitigungsrecht für die betriebliche Praxis,* **Loseblattsammlung, vols. 1 – 4,** WEKA Fachverlag für technische Führungskräfte, Augsburg 1994.
10. A. Hoschützky, H. Kreft: *Recht der Abfallwirtschaft,* vols. 1 and 2, Loseblattsammlung, Heymanns Verlag, Köln 1989.
11. G. Hösel, H. v. Lersner: *Recht der Abfallbeseitigung,* vols. 1 and 2, Loseblattsammlung, Erich Schmidt Verlag, Berlin 1972.
12. P. Kunig, G. Schwermer, L.-A. Versteyl: *Abfallgesetz,* 2nd ed., C. H. Beck'sche Verlagshandlung, München 1992.
13. H. v. Köller: *Leitfaden Abfallrecht,* 4th ed., Erich Schmidt Verlag, Berlin 1993.
14. S. Struß: *Abfallwirtschaftsrecht,* Heymanns Verlag, Köln 1991.
15. E. Edom: Ökonomische und öffentlich-rechtliche Probleme der Sonderabfallentsorgung, Erich Schmidt Verlag, Berlin 1991.

Specific References

16. Vom 10.4.1990 und vom 17.12.1990 (GMBl. S. 866) berichtigt am 4.2.1991 (GMBl. S. 136) und am 21.3.1991 (GMBl. S. 469); Gesamtfassung vom 12.3.1991 (GMBl. S. 139).
17. Vom 14. Mai 1993, BAnz. v. 29.05.1993, S. 4967.
18. BR-Drs. 594/92.
19. BverfG v. 26.11.1991 – 7 C 16/89 –.
20. Vom 3. April 1990, BGBl. I S. 648.
21. BVerwG Urt. v. 14.12.1973, NJW 1974 S. 815.
22. BGH Urt. v. 26. Februar 1991, NJW 1991, S. 1621.
23. Vgl. dazu Palandt, Bürgerliches Gesetzbuch, 53. Auflage 1994, Anm. 7 zu § 276.
24. BGBl. 1989 I S. 2198.
25. BGBl. 1990 I S. 2634.
26. Entsorga Schriften 14, Entsorga GmbH, Cologne, 1994, pp. 110 – 115.

27. Deutsches Institut für Normung: "Grundsätze produktbezogener Ökobilanzen – German 'Memorandum of Understanding'/'Conceptual Framework' (Stand Oct. 1993)," *DIN-Mitteilungen + elektronorm* **73** (1994) no. 3, 208 – 212.
28. Society of Environmental Toxicology and Chemistry (SETAC): Guidelines for LCA: A "Code of Practice" from Workshop held at Sesimbra, Portugal, Mar. 31 – Apr. 3, 1993.
29. Umweltbundesamt (eds.): Arbeitsgruppe Ökobilanzen: Ökobilanzen für Produkte, Bedeutung – Sachstand – Perspektiven, Berlin 1992.
30. R. Heijungs et al., *Environmental Life Cycle Assessment of Products, Guide & Backgrounds,* Centre of Environmental Science (CML), Leiden 1992.
31. G. Fleischer (ed.): "Produkt-UVP – Auswahl, Gestaltung und Optimierung umweltfreundlicher Produkte," in *Abfallverminderung,* EF-Verlag für Energie und Umwelttechnik, Berlin 1992.
32. W. Klöpffer, I. Renner: "Methodik der Wirkungsbilanz im Rahmen von Produkt-Ökobilanzen unter Berücksichtigung nicht oder nur schwer quantifizierbarer Umweltkategorien," UFO-Plan no. 100 01 102, Bericht der C.A.U. GmbH an das Umweltbundesamt Berlin, May 1994, erscheint in der Reihe UBA-Texte, Berlin 1994.
33. R. Grießhammer (ed.): *Produktlinienanalysen und Ökobilanzen,* Werkstattreihe, Öko-Institut, Freiburg 1991.
34. G. Fleischer (ed.): "Integrierter Technischer Umweltschutz – Wege zum umweltfreundlichen Werkstoff und zur ökologischen Betriebsoptimierung," in: *Produktionsintegrierter Umweltschutz,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994.
35. G. Fleischer: "Methodik des Vergleichs von Verwertung-/Entsorgungswegen im Rahmen der Ökobilanz," *Abfallwirtschafts Journal* **6** (1994) no. 10, 697 – 701.
36. G. Fleischer: "Der ökologische break-even-point für das Recycling," in K. J. Thomé-Kozmiensky: *Modelle für eine zukünftige Siedlungsabfallwirtschaft,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1993, pp. 95 – 108.
37. A. Braunschweig, R. Müller-Wenk: *Ökobilanzen für Unternehmungen. Eine Wegleitung für die Praxis,* Verlag Paul Haupt, Bern 1993, pp. 1 – 224.
38. G. Fleischer: "Eine schlüssige Ökobilanzmethode: Die Produkt-Umweltverträglichkeitsprüfung," *VDI Ber.* **1060** (1993) pp. 177 – 192.
39. G. Fleischer, W.-P. Schmidt: "LCA for New Composite Materials: Necessity and Methods of Considering Symmetry Rules," To be published in The International Journal of Life Cycle Assessment 1995.
40. SETAC: "Life-Cycle Assessment Data Quality: A Conceptual Framework," Workshop Report Society of Environmental Toxicology and Chemistry and SETAC Foundation for Environmental Education, Wintergreen, Virginia, Oct. 1992.
41. G. Fleischer, R. Ackermann, R. Schilling: "Ein Schritt auf dem Weg zur Ökobilanz," *Abfallwirtschafts Journal* **5** (1993) no. 5, 379 ff.
42. G. Fleischer, W.-P. Schmidt: "Voraussetzung zur Bestimmung des ökologisch optimalen Altpapieranteils mittels Ökobilanz: Nutzengleichheit," *Papier (Darmstadt)* **49** (1995) no. 6, pp. 334 – 347.
43. K.-H. Feuerherd: "Erhebung von Ökobilanzdaten in der Industrie am Beispiel des Projektes der europäischen Kunststoffherstellung," in: *GDCh* Fortbildungseminar Chemie, Seminar *"Die Ökobilanz – Eine Entscheidungshilfe"* Frankfurt/Main, Apr. 19, 1993.
44. G. Fleischer: "The Allocation of Open-Loop-Recycling in LCA," in: *Proceedings of the European Workshop on Allocation in LCA,* Centre of Environmental Science of Leiden University, Feb. 24 – 25, 1994, Leiden 1994, pp. 61 – 63.
45. F. Saykowski, M. Marsmann: "Stellenwert von Ökobilanzen im Bereich der Abfallwirtschaft," in: *VDI/GVC Jahrbuch 1995 – Verfahrenstechnik und Chemieingenieurwesen,* VDI-Verlag, Düsseldorf 1995.
46. U. H. A. de Haes: "Workshop Conclusions on Classification Session," in: *Workshop Report,* Congress on Dec. 2 – 3, 1991, Leiden, Apr. 1992, pp. 95 – 98.
47. J. Guniée, R. Heijungs: "Impact Assessment Within the Framework of Life Cycle Assessment of Products," in Vereinigung für ökologische Wirtschaftsforschung (eds.): *Informationsdienst des Institutes für ökologische Wirtschaftsforschung,* vol. 9, no. 1, Berlin 1994, pp. 7 – 9, A 20 178 f.

48. J. A. Assies: "State of Art in Lifecycle Assessment," in: *Workshop Report,* Congress on Dec. 2 – 3, Leiden, Apr. 1992, pp. 1 – 20.
49. G. Finnveden et al.: "Classificaton (Impact Analysis in Connection with LCA – A Preliminary Study," in Nordic Council of Ministers (eds.): *Product Life Cycle Assessment – Principles and Methodology,* Copenhagen – Stockholm 1992.
50. G. Fleischer: "Aus dem Leben einer Milchtüte . . . Die Bilanz des Abfalls," in *Funkkolleg Technik – einschätzen – beurteilen – bewerten,* Studieneinheit 10, Deutsches Institut für Fernstudienforschung an der Universität Tübingen, Beltz Verlag, Hemsbach 1994.
51. J. A. Annema: "Methodology for the Evaluation of Potential Action to Reduce the Environmental Impact of Chemical Substances," in: *Workshop Report,* Congress on Dec. 2 – 3, 1991, Leiden, Apr. 1992, pp. 73 – 80.
52. J. B. Guniée, R. Heijungs: "A Proposal for the Definition of Resource Equivalency Factors for Use in Product Life-Cycle Assessment," *Environ. Toxicol. Chem.* **14** (1995) no. 5, 917 – 925.
53. J. Guinée, R. Heijungs: "A Proposal for the Classification of Toxic Substances Within the Framework of Life Cycle Assessment of Products," *Chemosphere* **26** (1993) no. 10, 1925 – 1944.
54. Diskussionspapier des Münchner Kreises: Bewertung von Sach-Ökobilanzen – Zwischenbericht Jan. 1994, ILV München.
55. W.-P. Schmidt: "Life Cycle Assessment (LCA) zur Bestimmung der ökologisch optimalen Recyclingrate von Papier," Diplomarbeit, Technische Universität, Berlin 1994.
56. Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit (eds.): Umweltpolitik. Konferenz der Vereinten Nationen für Umwelt und Entwicklung im Juni 1992 in Rio de Janeiro. Dokumente. Klimakonvention. Konvention über die Biologische Vielfalt. Rio-Deklaration. Walderklärung. Bonn 1992.
57. Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit (eds.): Umweltpolitik. Konferenz der Vereinten Nationen für Umwelt und Entwicklung im Juni 1992 in Rio de Janeiro. Dokumente. Agenda 21. Bonn 1993.
58. K. Habersatter: "Ökobilanz von Packstoffen," *Schriftenreihe Umwelt des BUWAL,* no. 132, Bern 1990.
59. St. Ahbe, A. Braunschweig, R. Müller-Wenk: "Methodik für Ökobilanzen auf der Basis ökologischer Optimierung," *Schriftenreihe Umwelt des BUWAL,* no. 133, Bern 1990.
60. A. Bechmann: *Nutzwertanalyse, Bewertungstheorie und Planung,* 1st ed., Haupt-Verlag, Bern 1978.
61. J.-P. Hettelingh, R. J. Downing, P. A. M. de Smet: "Mapping Critical Loads for Europe," *CCE Technical Report,* no. 1, RIVM Report no. 259101001, Bilthoven, July 1991.
62. S. Schaltegger, A. Sturm: *Ökologieorientierte Entscheidungen in Unternehmen,* WWZ, Basel 1992.
63. Tellus Institute, Disposal Cost Fee Study, Boston 1991.
64. W. Klöpffer: "Environmental Hazard – Assessment of Chemicals and Products," part IV: Life Cycle Assessment (LCA), *Environ. Sci. Pollut. Res.* **1** (1994) no. 4, 271 – 279.
65. S. Schmitz, H.-J. Oelsen, A. Tiedemann: *Ökobilanz für Getränkeverpackungen,* part A: Methode der Berechnung und Bewertung von Ökobilanzen für Verpackungen, Umweltbundesamt, Berlin 1995.
66. J. B. Guinée: Data for the Normalization Step within LCA of Products, CML-Paper no. 14, Centre of Environmental Science, Leiden 1993.
67. Association of the Dutch Chemical Industry (VNCI): Integrated Substance Chain Management, Dec. 1991.
68. R. Brüggemann, B. Münzer, E. Halfon: "An Algebraic/graphical Tool to Compare Ecosystems with Respect to Their Pollution. The German River 'Elbe' as an Example," I: HASSE-Diagrams, *Chemosphere* **28** (1994) 863 – 872.
69. B. Grahl, J. Lohse, E. Schmincke: "Analyse methodischer Ansätze zur Klassifikation von Stoffen und Produkten aus dem Hausmüll in Hinblick auf deren Relevanz für die Vermeidung," im Auftrag des Büros für Technikfolgenabschätzung des deutschen Bundestages, Bonn 1992.
70. Bundesministerium für Bildung, Forschung, Wissenschaft und Technologie (eds.): *Deutscher Delphi-Bericht zur Entwicklung von Wissenschaft und Technik,* Institut für Systemtechnik und Innovationsforschung (Projektleitung), Bonn 1993.
71. A. Tiedemann: "Die ökologische Bewertung von Papier am Beispiel des Umweltzeichenentwurfs der Europäischen Gemeinschaft," Vortrag auf der UTECH Berlin 1994, Seminar Ökobilanzen.

72. Enquete-Kommission "Schutz des Menschen und der Umwelt" des Deutschen Bundestages (eds.): *Die Industriegesellschaft gestalten. Perspektiven für einen nachhaltigen Umgang mit Stoff- und Materialströmen,* Economica Verlag, Bonn 1994.
73. I. Boustead et al.: "Eco-Balance – Methodology for Commodity Thermoplastics," A Report for The European Centre for Plastics in the Environment (PWMI), Brussels 1992, p. 3.
74. F. Schmidt-Bleek: *Wieviel Umwelt braucht der Mensch?: MIPS – das Maß für ökologisches Wirtschaften,* Birkhäuser, Berlin 1993.
75. W. Gebler: "Ökobilanzen in der Abfallwirtschaft," *Stuttgarter Berichte zur Abfallwirtschaft,* **vol. 41,** Erich Schmidt Verlag, Berlin 1990.
76. Statistisches Bundesamt: *Umwelt,* Fachserie 19, Reihe 1.2: "Abfallbeseitigung im produzierenden Gewerbe und in Krankenhäusern," Wiesbaden 1990.
77. Statistisches Bundesamt: personal communication.
78. C. Christ: "Umweltschonende industrielle Arbeitsmethoden – Verfahrensverbesserungen und Stoffkreisläufe," in Bundesverband der Deutschen Industrie e. V.; *Industrieller Umweltschutz,* Tagungsband zum Seminar mit der Universität Leipzig, 1993, p. 210.
79. H. Hulpke, U. Müller-Eisen: "Abfallmanagement in der Chemischen Industrie," *Chem. Umwelt Tech.* 1993, 22.
80. C. Christ: "Konzepte zur Abfallverminderung und -vermeidung in der Chemischen Industrie," in H. Sutter, M. Held: *Stoffökologische Perspektiven der Abfallwirtschaft,* E. Schmidt Verlag, Berlin 1993, p. 83.
81. W. Hilger: *Chemie im Dialog "Herausforderung und Perspektiven,"* VCI-Dokumentation der Veranstaltung am 04.07.1991 im Wissenschaftszentrum Bonn, Frankfurt/Main, Oct. 1991.
82. H.-I. Joschek, G. Janisch: "Entsorgungskonzept eines Großbetriebes am Beispiel der BASF AG Ludwigshafen," in K. J. Thomé-Kozmiensky: *Sonderabfallwirtschaft,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1993.
83. K. Holzer: "Aerosolabscheidung aus Rauchgasen von Anlagen zur Verbrennung von Sonderabfällen," *Staub Reinhalt. Luft* **48** (1988) 203.
84. K. Capek: "Gemeinsame Verbrennung von Industrie-Klärschlamm und flüssigem Sonderabfall," *VGB Kraftwerkstech.* **70** (1990) 31.
85. Hoechst DE-OS 3 703 568 A 1, 1988 (G. Müller, H. Thomas, H. Schmolt).
86. B. Koglin, J. Beyer, R. Rink, J.-E. Roth: "Hochdruck-Verdichtung und Einbau von Abfällen, ein Systembestandteil der Bayer Kompakt-Deponie," *Chem. Ing. Tech.* **63** (1991) 605.
87. BASF, Umweltbericht 1993, Ludwigshafen.
88. Kali-Chemie, EP-A 0 474 093 A 1, 1991 (W. Legat et al.).
89. Kali-Chemie, EP-A 0 545 851 A 1, 1992 (H.-W. Swidersky et al.).
90. Kali-Chemie, EP-A 0 546 984 A 1, 1992 (H.-W. Swidersky et al.).
91. BASF, Pressegespräch zum rohstofflichen und werkstofflichen Recycling von Altkunststoffen bei der BASF, Ludwigshafen, March 16, 1994.
92. Hoechst, EP 0 212 410 B 1, 1990 (S. Hug et al.).
93. Verband der Chemischen Industrie e. V.: Pilotprojekt, Verwertung gebrauchter Industriepackmittel aus der Chemischen Industrie, Frankfurt/Main 1991.
94. Bayer, Umweltbericht 1993, Leverkusen.
95. Hoechst, Umweltschutz-Jahresbericht 1993, Frankfurt/Main.
96. G. Fleischer: "Strukturierung der Abfallvermeidung," in G. Fleischer (ed.): *Vermeidung und Verwertung von Abfällen 3,* EF-Verlag für Energie- und Umwelttechnik, Berlin, 1992.
97. W. R. Stahel: "Langlebigkeit der Produkte als Ausweg aus den Zielkonflikten beim Recycling?" in "Recyclinggerechte Produktentwicklung," *VDI-Ber.* 1089 (1993) 1 – 18.
98. W. Beitz: "Recyclinggerechtes Konstruieren," in: *TU International 20* **21** (1993) 14 – 15.
99. F. Fleischer: "Integrierter Technischer Umweltschutz – Wege zum umweltfreundlichen Werkstoff und zur ökologischen Betriebsoptimierung," in G. Fleischer (ed.): *Produktionsintegrierter Umweltschutz,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994, pp. 7 – 30.
100. S. Grieger, A. Wende: "Wege zur abfallvermeidenden Produktgestaltung," in G. Fleischer (ed.): *Produktionsintegrierter Umweltschutz,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994, pp. 91 – 124.

101. G. Fleischer: "Primäre Abfallvermeidung in der Industrie," in G. Fleischer (ed.): *Vermeidung und Verwertung von Abfällen 2,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1990, pp. 1 – 38.
102. W.-D. Mroß: "Produktionsintegrierter Umweltschutz: Vermeiden, Vermindern und Verwerten," in Zukunftsinitiative Rheinland-Pfalz (eds.): *Industrielle Ökosysteme – Vermeiden, Verwerten und Entsorgen in der Industriegesellschaft,* Herman Schmidt, Mainz 1994.
103. Dechema: "Produktionsintegrierter Umweltschutz – Verpflichtung der chemischen Industrie," Präambel in: *Produktionsintegrierter Umweltschutz in der chemischen Industrie,* Frankfurt/Main 1990, pp. 7 – 14.
104. J. A. Philipp, W. Theobald: "Herausforderungen des Umweltschutzes an den Werkstoff Stahl," in G. Fleischer (ed.): *Produktionsintegrierter Umweltschutz,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994, pp. 135 – 163.
105. Deutsches Institut für Normung: "Grundsätze produktbezogener Ökobilanzen – German "Memorandum of Understanding"/"Conceptual Framework" (Stand Oktober 1993)," *DIN-Mitteilungen + elektronorm* **73** (1994) no. 3, pp. 208 – 212.
106. Kreislaufwirtschafts- und Abfallgesetz – KrW-/AbfG (vom 27. 9. 1994).
107. G. Fleischer: "Der ökologische break-even-point für das Recycling," in K. J. Thomé-Kozmiensky: *Modelle für eine zukünftige Siedlungsabfallwirtschaft,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1993, pp. 95 – 108.
108. G. Fleischer, W.-P. Schmidt: "Voraussetzung zur Bestimmung des ökologisch optimalen Altpapieranteils mittels Ökobilanz: Nutzengleichheit", *Papier* (*Darmstadt)* **49** (1995) no. 6, pp. 344 – 347.
109. W. R. Stahel: *Langlebigkeit und Materialrecycling – Strategien zur Vermeidung von Abfällen im Bereich der Produktion,* Vulkan-Verlag, Essen 1991.
110. G. Fleischer: "The Allocation of Open-loop-Recycling in LCA," *Proceedings of the European Workshop on Allocation in LCA,* Leiden, Feb. 24 – 25, 1994, Brussels, pp. 61 – 63.
111. R. Steinhilper: "Industrielle Demontage von Serienprodukten – Praxis und Perspektiven," in: "Recycling, eine Herausforderung für den Konstrukteur," *VDI Ber.* **906** (1991) 103 – 125.
112. H. Wollenberger, B. Schenk: "Schaufeln und Scheiben in Gasturbinen – Werkstoff- und Bauteilverhalten," in *Forschung Aktuell,* nos. 39 – 41, Technische Universität, Berlin 1993, pp. 4 – 6.
113. K. H. Peters, H.-B. Lügen: "Stand der Roheisenerzeugung in der Bundesrepublik Deutschland," *Stahl Eisen* **110** (1990) no. 2, 55 – 64.
114. A. Wurm: "Umweltfreundlicher Güterverkehr – Systemtechnische Ansätze und Möglichkeiten," in G. Fleischer (ed.): *Vermeidung und Verwertung von Abfällen,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1989.
115. W. Bleck, W. Dahl, P. Hamacher: "Energieeinsparung durch endabmessungsnahe Herstellung von Stahlblechen," in G. Fleischer (ed.): *Produktionsintegrierter Umweltschutz,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994, pp. 165 – 180.
116. J. Vehlow, H. Vogg: "Thermische Zerstörung organischer Schadstoffe," in K. J. Thomé-Kozmiensky (ed.): *Müllverbrennung und Umwelt 5,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1991, pp. 447 –467.
117. W. R. Stahel: "Langlebigkeit von Gütern und Material-Recycling – Abfallvermeidung und Ressourceneinsparungen in der Nutzung," in G. Fleischer (ed.): *Vermeidung und Verwertung von Abfällen 3,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1992, pp. 329 – 343.
118. M. Socher: "Integrierte Werkstofforschung – Perspektiven und Konsequenzen," in G. Fleischer (ed.): *Produktionsintegrierter Umweltschutz,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994, pp. 55 – 90.
119. H. Käufer, M. Grimm: "Recyclinggerechte Produktgestaltung durch Kunststoffanwendungen in neuen Dimensionen," in: *Recycling – eine Herausforderung für den Konstrukteur,* Tagung Bad Soden, Nov. 14.– 15. 1991, VDI-Verlag, Düsseldorf 1991.
120. H. Tober: "Recyclinggerechtheit von "Weißer Ware" – Closed-Loop-Engineering im Team," in: "Recyclinggerechte Produktentwicklung," *VDI Ber.* **1089** (1993) 141 – 156.
121. W. Jorden: "Konstruieren recyclinggerechter Produkte mit der neuen Richtlinie VDI 2243," in: "Recycling, eine Herausforderung für den Konstrukteur," *VDI Ber.* **906** (1991) 23 – 41.

122. P. Schulz: "Recyclinggerechte Konstruktion von Fernsehgeräten," in: "Recyclinggerechte Produktentwicklung," *VDI Ber.* **1089** (1993) 63 – 74.
123. M. Schmidt-Kretschmer, W. Beitz: "Demontagefreundliche Verbindungstechnik – ein Beitrag zum Produktrecycling," in: "Recycling, eine Herausforderung für den Konstrukteur," *VDI Ber.* **906** (1991) 153 – 170.
124. J. Barrenscheen, M. Renken: "Recyclinggerechte Verbindungstechnik im Automobil," in: "Recyclinggerechte Produktentwicklung," *VDI Ber.* **1089** (1993) 47 – 61.
125. A. Weber: *Neue Werkstoffe,* VDI-Verlag, Düsseldorf 1989.
126. Manufacture of Iron and Steel, A.D. 1857 Nr. 221, Letters Patent to Henry Bessemer, of Queen Street Place, New Cannon Street, in the City of London, for the Invention of "Improvements in the Manufacture of Iron and Steel," Sealed the 3rd April 1857, and dated the 24th January 1857.
127. F. K. Iverson, K. Busse: "Ein Jahr Betriebserfahrung mit der CSP-Anlage für Warmband bei Nucor Steel," *Stahl Eisen* **111** (1991) no. 1, 37 – 45.
128. G. Spur, T. Stöferle (eds.): *Urformen; Handbuch der Fertigungstechnik,* **vol. 1,** Hanser Verlag, München – Wien 1981, pp. 951 – 967.
129. L. Deininger: "Der QSL-Prozeß zur Bleigewinnung," in G. Fleischer (ed.): *Vermeidung und Verwertung von Abfällen 3,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1992, pp. 471 – 503.
130. H. Schaum: "Vermeidung und Verwertung von Sonderabfällen in der chemischen Industrie," in K. J. Thomé-Kozmiensky (ed.): *Behandlung von Sonderabfällen 2,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1988, pp. 716 – 727.
131. J. Weitkamp: "Katalyse in Zeolithen – neue Perspektiven für die präparative Chemie," in: Gesellschaft Deutscher Chemiker (GDCh) (eds.): 22. Hauptversammlung, Bonn, Sept. 18 – 22, 1989, Programm und Kurzreferate, VCH, Weinheim 1989, p. 105.
132. E. Winterfeldt: "Effizienz und Selektivität – Beiträge zur umweltschonenden Synthesechemie," in: Gesellschaft Deutscher Chemiker (GDCh) (eds.): 22. Hauptversammlung, Bonn, Sept. 18 – 22, 1989, Programm und Kurzreferate, VCH, Weinheim 1989, p. 109.
133. G. Fleischer, R. Kammel, U. Lembke: "Verarbeitung von flüssigen Kupferlegierungsmaterial im Altmetallkonverter," *Metall* **32** (Berlin) (1978) no. 2, 129 – 134.
134. K. Fischwasser, H.-W. Lieber, J. Müller: "Vermeidung von Abwasser und nicht verwertbaren Rückständen durch Regenerierung galvanischer Elektrolyte," *Galvanotechnik* **85** (1994) no. 7, pp. 2294 – 2301.
135. H.-W. Lieber: "Abwasser- und Abfallvermeidung in der Metallbe- und -verarbeitenden Industrie," *Gewässerschutz Wasser Abwasser* **135** (1993), pp. 261 – 274.
136. H. Schene, H. Vesper: "Schadstoffarme industrielle Lackiertechnik," in G. Fleischer (ed.): *Vermeidung und Verwertung von Abfällen,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1989.
137. H. Sutter: "Vermeidung und Verwertung von Sonderabfällen," Reihe: *Abfallwirtschaft in Forschung und Praxis,* vol. 17, Erich Schmidt Verlag, Berlin 1987, pp. 64 – 66.
138. H. Sutter, B. Mahrwald, M. Grosse Ophoff: "Die Bedeutung der ökologischen Rangfolge im neuen Kreislaufwirtschafts- und Abfallgesetz," Umwelt Technologie Aktuell**5** (1994) no. 6, 427 –436.
139. G. Fleischer: "Aus dem Leben einer Milchtüte . . . Die Bilanz des Abfalls," in Deutsches Institut für Fernstudienforschung an der Universität Tübingen (eds.): *Funkkolleg Technik: einschätzen – beurteilen – bewerten,* Studieneinheit 10, Tübingen 1994.
140. Society of Environmental Toxicology and Chemistry (eds.): "Guidelines for Life-Cycle Assessment: A "Code of Practice"," 1st ed., *SETAC-Workshop,* Sesimbra, Portugal, Mar. 31 – Apr. 3, 1993, Brüssel, Pensacola: SETAC 1993.
141. Kreislaufwirtschafts- und Abfallgesetz – KrW-/AbfG (vom 27.9.1994).
142. VDI-Richtlinie 2243 (Entwurf): Konstruieren recyclinggerechter technischer Produkte, VDI, Düsseldorf 1991.
143. E. Höffgen, W. Ullrich, H. Schicks: "Werkstoff Stahl: Unentbehrlich für eine Industriegesellschaft, umweltfreundlich und immer wieder recyclebar," in G. Fleischer (ed.): *Vermeidung und Verwertung von Abfällen,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1992, pp. 615 – 629.

144. W. D. Riley: "Methodologies to Identify Scrap Metals," in J. H. L. van Linden, D. L. Steward, Y. Sahai (eds.): *Second International Symposium – Recycling of Metals and Engineered Materials,* TMS, New York 1990, pp. 575 – 583.
145. R. D. Prengaman: "Recycling of Lead," in: *Metals Handbook,* vol. 2, ASM, London, USA 1990, pp. 1221 – 1223.
146. W. Nieder: "Das Totalrecycling von Altfahrzeugen," *Radex Rundsch.* 1993, nos. 1/2, 308 – 313.
147. G. Fleischer: "Voraussetzungen und Grenzen des stofflich sinnvollen Recyclings," *Radex Rundsch.* 1993, nos. 1/2, 274 – 278.
148. T. Gabler: *Wirtschafts-Lexikon,* Betriebswirtschaftlicher Verlag, Dr. Th. Gabler, Wiesbaden 1991.
149. G. Fleischer: "Der ökologische Break-even-point für das Recycling," in K. J. Thomé-Kozmiensky: *Modelle für eine zukünftige Siedlungsabfallwirtschaft,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1993, 95 – 108.
150. G. Fleischer: "Methodik des Vergleichs von Verwertung-/Entsorgungswegen im Rahmen der Ökobilanz," *AbfallwirtschaftsJournal* 1994, no. 10, 697 –701.
151. R. Scholz: Gutachten "Verfassungsfragen zur Entsorgung von Kunststoffabfällen nach dem Abfallgesetz vom 27.8.1986 und der Verpackungsverordnung vom 12.6.1991," KJ-Verlag, Bad Homburg 1990.
152. B. Böndel: "Zur Entsorgung wieder ins Fahrzeug zurück," *VDI Nachr.* **44** (1990) no. 9, 24.
153. G. Walter, P. Eyrer: "Zum Wegwerfen zu schade," Recyclingkonzepte der Mercedes Benz AG, Daimler Benz AG, Untertürkheim 1989.
154. J. Siebenlist: "Aus dem Auto wieder ins Auto," *VDI Nachr.* **44** (1990) no. 28, 2.
155. K. Wittig: Diskussionsbeitrag anläßlich des 15. kunststofftechnischen Kolloquiums des IKV, Aachen, March 1990.
156. A. G. Haardt: "Polypropylen-Schaumstoff, energieabsorbierender Werkstoff für Stoßfängerkerne," *Kunststoffe* **78** (1989) no. 3, 256 – 259.
157. G. Menges, H. Emminger, V. Lackner: "Recycling of Plastics," Vortrag auf der MATtech 90, Helsinki, June 1990.
158. B. Aur: "Nur Sortenreinheit verspricht Zweitnutzen," *VDI Nachr.* **44** (1990) nos. 12/13, 31.
159. M. Gebauer, K. Bühler: "Recycling von Kunststoffen," Das Verwertungskonzept der Leuna-Werke AG, part 2, 1992.
160. Verband Kunststofferzeugende Industrie: Forschungsprogramm Wiederverwertung von Kunststoffabfällen, Zusammenfassung der Teilprojekte 1 – 8, Frankfurt 1981.
161. s. [150]
162. G. Bauer: "Recycling von Polyurethanabfällen und gemischten Polymerabfällen durch Alkoholyse," in: K. J. Thomé-Kozmiensky (ed.): *Recycling international,* EF-Verlag, Berlin 1989, p. 293 ff.
163. M. Baerns et al.: Gutachten "Stofflicher und energetischer Wirkungsgrad der hydrierenden Aufarbeitung von Altkunststoffen nach dem Verfahren der Kohleöl-Anlage Bottrop, Universität Bochum, Lehrstuhl für technische Chemie, June 1993.
164. s. [150]
165. B. Bilitewski, G. Härdtle, K. Marek: "Systemvergleich über Verfahren zur Pyrolyse von Abfällen," unveröffentlichter Bericht des Ingenieurbüros für Technischen Umweltschutz, Berlin, Halberstädterstrasse 7, Jan. 1985.
166. R. Schmidt: *Müllpyrolyse,* Vortrag im VDI-Bildungswerk, BW 7749.
167. G. Menges, R. Fischer, V. Lackner, EP 9 100 959, 1990. G. Menges, V. Lackner: "Gaserzeugung aus Kunststoffabfällen," in G. Menges, W. Michaeli, M. Bittner (eds.): *Recycling von Kunststoffen,* Hanser Verlag, München-Wien 1992.
168. B. Buttker, W. Rabe, J. Müller: Kurzbericht über erste Erfahrungen der Verwertung von Reststoffen in der Festbettdruckvergasung in Schwarze Pumpen, Seminar des VDI-Bildungswerkes "Verbrennungsanlagen für Sonderabfall", Mannheim, Jan. 27 – 28, 1993.
169. G. Menges, R. Fischer: "Kohlenstoff-Recycling beim Aufarbeiten gemischter Kunststoffabfälle," *Kunststoffe* **81** (1991) no. 1, 6 – 14.
170. Zukunftsenergien – Fakten und Argumente, Veba AG, Essen 1989.
171. G. Fischer, H. G. Kitzerow: "Gedanken zur Zukunft der thermischen Haus- und Gewerbemüll-Entsorgung," *Abfallwirtsch. J.* **1** (1989) nos. 7/8, 16 – 19.
172. E. Dierks: "Thermische Verwertung von Kunststoffen in Haushaltsabfällen," in: G. Menges, W. Mi- chaeli, M. Bittner (eds.):

Recycling von Kunststoffen, Hanser Verlag, München-Wien 1992.
173. *KJ-Mitteilungen* **4** (1987) no. 2, 1.
174. W. Biendarra: "Bericht über eine Untersuchung zum Energierecycling der Leichtfraktion eines Automobilshredders in einem Drehrohrofen," Institut für gewerbliche Wasserwirtschaft und Luftreinhaltung e. V., Köln 1987.
175. K. Weirauch: "Kunststoffentsorgung – Recycling," interner Bericht der Bayer AG, Leverkusen, Feb. 20, 1989.
176. K. D. Kohnle: "Thermische Verwertung von Restkunststoffen – Übersicht über mögliche Technologien," Vortrag gehalten auf dem 10. Leobener Kunststoffkolloquium, May 10 – 11, 1990.
177. R. Fischer, G. Menges, EP-OS 00 927, 1988.
178. B. E. Naumann, C. J. Lynch; "Renselaer's Selective Dissolution Process", *KJ-Information,* 30 222.
179. Gesetz über die Vermeidung und Entsorgung von Abfällen (Abfallgesetz – AbfG) vom 27.8.1986 (BGBl. I S. 1410, ber. S. 1501), zuletzt geändert am 26.6.1992 (BGBl. I S. 1161)
180. Gesetz zum Schutz vor schädlichen Umwelteinwirkungen durch Luftverunreinigungen, Geräusche, Erschütterungen und ähnliche Vorgänge (BundesImmissionsschutzGesetz – BImSchG) in der Fassung der Bekanntmachung vom 14.5.1990 (BGBl. I S. 880), zuletzt geändert am 1.5.1993
181. Verordnung über die Entsorgung gebrauchter halogenierter Lösemittel (HKWAbfV) vom 23.10.1989 (BGBl I S. 1918)
182. W. A. Gruse, D. R. Stevens: *Chemical Technology of Petroleum,* McGraw-Hill Book Company, New York-Toronto-London 1960.
183. F. Asinger: *Die Petrolchemische Industrie,* part I and II, Akademie-Verlag, Berlin 1971.
184. B. Riediger: *Die Verarbeitung des Erdöles,* Springer Verlag, Berlin-Heidelberg-New York 1971.
185. M. Bank: *Basiswissen Umwelttechnik – Wasser, Luft, Abfall, Lärm, Umweltrecht,* Vogel Buchverlag, Würzburg 1993.
186. Ingenieurgesellschaft für Technischen Umweltschutz GmbH Berlin: Entsorgungsstruktur und Lösungen auf dem Gebiet der Sonderabfälle in den Wirtschaftszweigen Hüttenindustrie, Chemische Industrie, Metallbe- und -verarbeitung (Großmengenkonzept), Studie im Auftrag der Freien und Hansestadt Hamburg – Umweltbehörde, 1989.
187. W. R. A. Vauck, H. A. Müller: *Grundoperationen chemischer Verfahrenstechnik,* Deutscher Verlag für Grundstoffindustrie, Leipzig 1981.
188. H. Schubert: *Aufbereitung fester mineralischer Rohstoffe,* vols. I – III, Deutscher Verlag für Grundstoffindustrie, Leipzig 1989.
189. W. Drechsler: "Abfallverbrennung auf der Nordsee – Konzepte zur Vermeidung und Verwertung von CKW-Abfällen," *EntsorgungsPraxis* **3** (1989) 82 – 86.
190. J. Schulze: "Entwicklungstendenzen zu einer rückstandsfreien oder rückstandsarmen Chemieproduktion," *Müll Abfall* **19** (1987) no. 1, 14 ff.
191. Hoechst, Verwertungsmodell Hoechst für FKW-Kältemittel, publication no. KT, 02/87, Frankfurt/Main.
192. H. Krill: "Kontinuierliche Prozesse wachsen langsam," *Chem. Ind.* **111** Düsseldorf (1988) no. 2, 76 – 78.
193. A. Gorak, L. Nowicki: "Rückgewinnung von organischen Lösungsmitteln aus verdünnten Gemi- schen, " in K. J. Thomé-Kozmiensky (ed.): *Behandlung von Sonderabfällen,* EF-Verlag, Berlin 1987.
194. M. Izumo: "Ein Verfahren zur Desodorierung und Wiederverwertung von Lösungsmitteln," in K. J. Thomé-Kozmiensky (ed.): *Behandlung von Sonderabfällen,* EF-Verlag, Berlin 1987.
195. "Abluftreinigung," *Verfahrenstechnik* **27** (1993) nos. 1 – 2, 34.
196. UTZ Umwelttechnisches Zentrum: "Entfernung und Reinigung von Oberflächen – Chlorkohlenwasserstoffe, Benzinkohlenwasserstoffe oder wäßrige Systeme," UTZ-Materialien no. 9, Frankfurt 1989.
197. L. Weimann: "Wann welche Methoden anwenden bei der Lösungsmittelrückgewinnung?", Maschinenmarkt (Würzburg) **80** (1974) no. 11, 298 – 300.
198. H. Sutter: *Beseitigung, Vermeidung und Verwertung von Lackschlämmen,* Müll-Handbuch, Kennzahl 8551, Erich Schmidt Verlag, Berlin 1986.
199. Ingenieurgesellschaft für Technischen Umweltschutz GmbH Berlin: Bericht über die Untersuchung von 8 Lackier- und Beschichtungsanlagen im Land Berlin, Studie im Auftrag der Senatsentwicklung für

Stadtentwicklung und Umweltschutz, Berlin 1993.
200. K.-J. Fischer, J. Niestroj, H. Alwast, B. Bilitewski: "Die Umsetzung des Reststoffvermeidungs- und -verwertungsgebotes nach § 5 Abs. 1 Satz 3 BImSchG am Beispiel von Lackieranlagen," *Abfallwirtsch. J.* **3** (1991) no. 11, 741 – 745.
201. B. Bilitewski, G. Härdtle, K. Marek: *Abfallwirtschaft – Eine Einführung,* 2nd ed., Springer Verlag, Berlin-Heidelberg-New York 1994.
202. I. Hanhoff Stemping: "Anforderungen an Lackieranlagen nach BImSchG," *Oberfläche + JOT* (1992) no. 5, 48 – 54.
203. Gesetz zum Schutz vor schädlichen Umwelteinwirkungen durch Luftverunreinigungen, Geräusche, Erschütterungen und ähnliche Vorgänge (BundesImmissionsschutzGesetz – BImSchG) in der Fassung der Bekanntmachung vom 14.5.1990 (BGBl. I S. 880), zuletzt geändert am 1.5.1993
204. H. Vesper: "Vermeidung und Verwertung von Lackoverspray," in: "Umweltschonende und abfallarme Lackiertechniken," *Workshop im Rahmen des umweltpolitischen Kongresses der Landesregierung Baden-Württemberg,* 1991.
205. H. Sutter: "TA-Abfall, Auswirkungen auf Lackiertechnik," in Deutsche Forschungsgesellschaft für Oberflächenbehandlung e. V. (eds.): *Neue Entwicklungen in der Lackiertechnik,* part 2: "Lackwiederverwendung und Emissionen", Conference, Hannover 1987.
206. Fa. Eisenmann: *Leitfaden für den Lackierbetrieb,* 24th revised ed., Böblingen 1990.
207. "Lösungen für das Lack- und Lösemittelrecycling," reprint of *Zukunft heute,* Rhein-Neckar ed., no. 1, 1991.
208. J. Demmrich, E. Weissflog, U. Nagel, A. Nahmer: "Aufbereitung und Verwertung von Lackkoagulaten," *Chemie-Anlagen und Verfahren (cav)* **23** (1990) 12.
209. VDI-Richtlinie 2243 (draft), Konstruieren recyclinggerechter technischer Produkte, VDI, Düsseldorf 1991.
210. *Winnacker-Küchler,* 4th ed., **4,** pp. 554 – 559.
211. *Winnacker-Küchler,* 4th ed., **4,** p. 467.
212. J. Schiebisch, G. W. Ehrenstein: "Teil 7/9 Eisen und Stahl, Teil 7/10 Nichteisenmetalle, Teil 7/10.1 Kupfer, Teil7/10.2 Aluminium," in Th. Brinkmann, G. W. Ehrenstein, R. Steinhilper (eds.): *Umwelt- und recyclinggerechte Produktentwicklung: Anforderungen, Werkstoffwahl, Gestaltung, Praxisbeispiele,* WEKA- Fachverlag für Technische Führungskräfte (Losebl.-Ausg.), Augsburg 1994.
213. L. D. Redden, R. Swarup: "Chemical Processing of Hardface Alloy Grinding Waste," in J. H. L. van Linden, D. L. Stewart, Y. Sahai (eds.), *Second International Symposium – Recycling of Metals and Engineered Materials, TMS* 1990, 171 – 173.
214. J. L. Holman, L. A. Neumeier: "Reclamation of Metals from Magnet Alloy Grinding Sludge," in Metallurgical Society of AIME (eds.): *International Symposium on Recycle and Secondary Recovery of Metals,* Fort Lauderdale, FL 1985, pp. 327 – 335.
215. U. Strenge: "Vermeidung und Verwertung von Beizsäuren," in H. Sutter (ed.): *Vermeidung und Verwertung von Abfällen,* EF-Verlag für Umwelttechnik, Berlin 1989, pp. 177 – 187.
216. E. Gock, J. Kähler: "Naßchemische Verfahren in der Abfallaufbereitung," in K. J. Thomé-Kozmiensky (ed.): *Materialrecycling durch Abfallaufbereitung,* EF-Verlag für Umwelttechnik, Berlin 1992, p. 408.
217. R. R. Odle, I. Martinez, L. A. Deets: "The Recycling of Hazardous Metal Plating Wastes," *JOM* **43** (1991) no. 6, 28 – 31.
218. K. Bosse: "Verminderung oder Verwertung von Galvanikrückständen," *Entsorgungspraxis* **12** (1994) no. 4, 66.
219. H. Maczek: "Möglichkeiten des Recycling von Zinkabfällen im IS-Ofen," in K. J. Thomé-Kozmiensky (ed.): *Recycling International: Gewinnung von Energie und Material aus Rückständen und Abfällen,* E. F. – Verlag für Umwelttechnik, Berlin 1982, pp. 759 –764.
220. H. Maczek, W. Massion: "Umwelt, Energie und Recycling," *Erzmetall* **44** (1991) no. 12, 585 – 596.
221. W. Mehner: "Die Verhüttung von Zinkerz und Zink-Sekundärvorstoffen nach dem New Jersey-Verfahren im Hüttenwerk Harz," *Erzmetall* **38** (1985) no. 10, 493.
222. P. Isenrath: "Recycling von kupferhaltigen Materialien – Stand der Technik und Entwicklungstendenzen," *Metall (Berlin)* **47** (1993) no. 2, 184.
223. R. Steinhilper, U. Hudelmaier: *Erfolgreiches Produktrecycling zur erneuten Verwendung oder Verwertung. Ein Leitfaden für Unternehmer,* Fraunhofer-Institut für

Produktionstechnik und Automatisierung, Stuttgart 1993, p. 203.
224. R. Steinhilper: "Teil 9/2 Auswahl der Verfahren, 9/2.1 Verfahrensschritte und Verfahrensbestimmung, 9/2.2 Branchenbezogene Auswahl," in Th. Brinkmann, G. W. Ehrenstein, R. Steinhilper (eds.): *Umwelt und recyclinggerechte Produktentwicklung: Anforderungen, Werkstoffwahl, Gestaltung, Praxisbeispiele,* WEKA-Fachverlag für Technische Führungskräfte (Losebl.-Ausg.), Augsburg 1994.
225. DIN 8580,Fertigungsverfahren, Einteilung, Beuth Verlag, Berlin 1974.
226. K. J. Thomé-Kozmiensky (ed.): *Kreislaufwirtschaft,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994, pp. I/285 – I/320.
227. E. Höffken, W. Ullrich, H. Schicks: Werkstoff Stahl: "Unentbehrlich für eine Industriegesellschaft, umweltfreundlich und immer wieder recyclebar, " in G. Fleischer (ed.): *Vermeidung und Verwertung von Abfällen,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1992, pp. 615 – 629.
228. W. Nieder: "Das Totalrecycling von Altfahrzeugen," *Radex-Rundsch.* 1993, nos. 1/2, 308 – 313.
229. K. J. Thomé-Kozmiensky (ed.): *Kreislaufwirtschaft,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994, pp. I/285 – I/320.
230. A. Brumby: "Der Kreislauf der Edelmetalle bei Autoabgaskatalysatoren," *Metall (Berlin)* **42** (1988) no. 7, 711 – 713.
231. J. E. Hoffmann: Recovering Platinum-Group Metals from Auto Catalysts, *J. Met.* **40** (1988) no. 6, 40 – 45.
232. P. N. H. Bhakta: "Recent Technology and Trends in Automobile Recycling," *JOM* **46** (1994) no. 2, 36 – 39.
233. G. Fleischer, M. Bargfrede, U. Schiller: *Vermeidung und Verwertung von Abfallsäuren und Chemiegipsen,* Umweltbundesamt, Berlin 1994, 12-8 – 12-44.
234. G. Schenker: "Bleirecycling aus Akkumulatorenschrott und anderen Sekundärvorstoffen in der Bleihütte Oker der Metaleurop," *Abfallwirtsch. J.* **3** (1991) no. 1/2, 30 – 55.
235. R. D. Prengaman: "Recycling of Lead," in: *Metals Handbook,* vol. 2, ASM, Metals Park, Ohio 1990, pp. 1221 – 1223.
236. K. J. Thomé-Kozmiensky (ed.): *Materialrecycling durch Abfallaufbereitung,* EF-Verlag für Umwelttechnik, Berlin 1992, pp. 207 – 212.
237. J. Hanulik, J. F. Goldinger, J. Abraham, W. Plieth: "Recycling von Trockenbatterien," in G. Fleischer (ed.): *Abfallvermeidung in der Metallindustrie,* EF-Verlag für Umwelttechnik, Berlin 1989, pp. 281 – 292.
238. K. J. Thomé-Kozmiensky (ed.): *Materialrecycling durch Abfallaufbereitung,* EF-Verlag für Umwelttechnik, Berlin 1992, pp. 184 – 185.
239. K. J. Thomé-Kozmiensky (ed.): *Materialrecycling durch Abfallaufbereitung,* EF-Verlag für Umwelttechnik, Berlin 1992, pp. 200 – 201.
240. B. E. Langer: "Recycling von Elektronikschrott," *Metall (Berlin)* **48** (1994) no. 11, 880 – 885.
241. J. E. Hoffmann: "Recovering Precious Metals from Electronic Scrap," *JOM* **44** (1992) no. 7, 43 – 48.
242. R. Steinhilper, M. Hieber: "Entwicklung eines umfassenden Recyclingkonzepts für die Computerindustrie," in G. Fleischer (ed.): *EDV, Elektronikschrott, Abfallwirtschaft,* EF-Verlag für Umwelttechnik, Berlin 1993.
243. W. Thiele, H. Matschiner: "Neue Wege zur Aufarbeitung von Elektronik-Schrott," in G. Fleischer (ed.): *Elektronikschrott, Abfallwirtschaft,* EF-Verlag für Umwelttechnik, Berlin 1993, pp. 193 – 203.
244. J. M. Larrain, W. D. Riley, R. D. Brown: "Spark Testing in the Identification of Stainless Steel and Superalloy Scrap," in Metallurgical Society of AIME (eds.): *International Symposium on Recycle and Secondary Recovery of Metals,* Fort Lauderdale, FL 1985, pp. 365 – 379.
245. W. D. Riley: "Methodologies to Identify Scrap Metals," in J. H. L. van Linden, D. L. Stewart, Y. Sahai (eds.): *Second International Symposium – Recycling of Metals and Engineered Materials, TMS 1990,* 575 – 583.
246. K. Sczimarowsky: "Kontinuierliche computergesteuerte Metall-Legierungs-Sortierung," in: G. Fleischer (ed.): *Abfallvermeidung in der Metallindustrie,* EF-Verlag für Umwelttechnik, Berlin 1989, pp. 107 – 117.
247. H.-P. Sattler: "Automatic Sorting of Non-Ferrous Metals from Automobile Shredders," in J. H. L. van Linden, D. L. Stewart, Y. Sahai (eds.): *Second International*

Symposium – Recycling of Metals and Engineered Materials, TMS **1990,** 333 – 341.
248. L.-P. Pape: "Untersuchungen zur Optimierung des Aluminiumkreislaufs," Dissertation, Technische Universität Berlin 1995, pp. 12 – 13.
249. H. A. Faure: "Entwicklung, Stand der Technik und Zukunftsaspekte der Stahlerzeugung," *Stahl Eisen* **113** (1993) no. 6, 39.
250. R. Berlekamp, H. W. Kreutzer, R. Willeke: "Schrott – ein weltweit gefragter Rohstoff für die Stahlherstellung," *Stahl Eisen* **109** (1989) no. 22, 1073.
251. J. A. Phillip et al.: "Recycling in der Stahlindustrie," *Stahl Eisen* **112** (1992) no. 12, 75.
252. H. W. Kreutzer: "Vom Schrott zum Stahl," *Stahl Eisen* **110** (1990) no. 6, 129.
253. K. Giepen: "Aluminium-Recyclingofen mit neuer Technik," *Aluminium (Düsseldorf)* **66** (1990) no. 5, 438.
254. W. Ullrich, H. Schicks: "Aspekte zum Recycling von metallisch beschichtetem Stahl," *Stahl Eisen* **111** (1991) no. 11, 85.
255. Bundesverband der Deutschen Schrottwirtschaft e. V. (BDS) (eds.): *Vom Schrott zum Stahl,* Verlag Handelsblatt, Düsseldorf 1984.
256. K. Krone et al.: "Ökologische Aspekte der Primär- und Sekundäraluminiumerzeugung in der Bundesrepublik Deutschland," *Metall (Berlin)* **44** (1990) no. 6, 559.
257. E. L. Rooy, J. H. L. Van Linden: "Recycling of Aluminum," in *Metals Handbook,* vol. 2, ASM, Metals Park, OH, 1990, pp. 1206 – 1213.
258. G. Fleischer, R. Kammel, U. Lembke: "Verarbeitung von flüssigem Kupferlegierungsmaterial im Altmetallkonverter," *Metall (Berlin)* **32** (1978) 129 – 135.
259. G. Fleischer, R. Kammel, U. Lembke: "Zur pyrometallurgischen Verarbeitung von zinn-, blei- und zinkhaltigen Flugstäuben," in K. J. Thomé-Kozmiensky (ed.): *Recycling von Metallen,* EF-Verlag für Umwelttechnik, Berlin 1987, pp. 94 – 95.
260. R. Boecker: "Betrachtungen zum Stand des Metall- und Kunststoffrecyclings," *Erzmetall* **45** (1992) no. 10, 531 – 539.
261. C. O. Bounds: "Recycling of Zinc from EAF Dust," in *Metals Handbook,* vol. 2, ASM, Metals Park, OH, 1990, pp. 1224 – 1226.
262. H. E. Vest: "Vakuumdestillation von zinkhaltigen metallischen Rest- und Abfallstoffen," Dissertation, Rheinisch-Westfälische Technische Hochschule, Aachen 1991.
263. M. Beckmann: "Aufarbeitung von Aluminiumsalzschlacken in Nordrhein-Westfalen," *Aluminium (Düsseldorf)* **67** (1991) no. 6, 586.
264. M. Beckmann: "Aufarbeitung von Aluminiumsalzschlacken in Nordrhein-Westfalen," in E. Keller, W. Schenkel (eds.): *Abfallwirtschaft und Recycling,* Vulkan-Verlag, Essen 1992, pp. 51 – 54.
265. Klärschlammverordnung (AbfKlärV), Bundesgesetzblatt, part 1 (1992).
266. Dritte Allgemeine Verwaltungsvorschrift zum Abfallgesetz, TA-Siedlungsabfall (1993).
267. U. Möller: "Klärschlamm, Produkt der Abwasserbelastung und der Abwasserreinigung," in: *ATV-Handbuch,* vol. 9, Klärschlamm, Ernst & Sohn, Berlin 1995, chap. 1.2.
268. ATV/BDE/VKS – Fachausschuß 3.2, Arbeitsbericht "Maschinelle Schlammentwässerung," *Korresp. Abwasser* **42** (1995) 271 – 285.
269. U. Loll: "Aerobe Stabilisation," in: *ATV-Handbuch,* vol. 9, Klärschlamm, Ernst & Sohn, Berlin 1995, chap. 4.2.
270. U. Loll: "Duale biologische Stabilisierung," in: *ATV-Handbuch,* vol. 9, Klärschlamm, Ernst & Sohn, Berlin 1995, chap. 4.4.
271. ATV/VKS-AG 3.2.2, Entseuchung von Klärschlamm: 1. Arbeitsbericht, *Korresp. Abwasser* **33** (1986) 1141 – 1142; 2. Arbeitsbericht, *Korresp. Abwasser* **35** (1988) 71 – 74; 3. Arbeitsbericht, *Korresp. Abwasser* **35** (1988) 1325 – 1333.
272. K. J. Thomé-Kozmiensky (ed.): *Behandlung von Sonderabfällen 3,* EF-Verlag für Energie- und Umwelttechnik GmbH, Berlin 1990.
273. GVC-Tagung: Entsorgung von Sonderabfällen durch Verbrennung, Baden-Baden, Dec. 1989.
274. Gesamtfassung der Zweiten allgemeinen Verwaltungsvorschrift zum Abfallgesetz (TA Abfall), geltende Fassung vom 01. April 1991.
275. Hessische Landesanstalt für Umwelt: Merkblatt Technische Bestimmungen für Sonderabfall-Kleinmengen-Zwischenläger, no. 128, 1991.
276. H. Sutter (ed.): *Industrielle Sonderabfälle und Reststoffe: Vermeidung und Verwertung in Sonderabfall, Leitfaden für Kommunen, Wirtschaft und Politik,* Verlag Bonner Energie Report, Bonn 1990.

277. 2. Allgemeine Verwaltungsvorschrift zum Abfallgesetz (TA-Abfall), Bundesanzeiger Verlags-Ges., Köln 1991.
278. Länderarbeitsgemeinschaft Abfall in Hösel/Schenkel/Schnurer (eds.) "LAGA-Informationsschrift Abfallarten," in: *Müllhandbuch,* vol. 2, Erich Schmidt Verlag, Berlin 1990.
279. F. Schoder: "Die Eindampfung von Sickerwasser aus Sonderabfall-Deponien und anderer problematischer Abwässer; Verfahrensbeschreibungen und Betriebserfahrungen" in *Tagung Deponiesickerwasserreinigung,* München, April 16, 1991.
280. J. Kollbach et al.: "Deponiesickerwasserreinigungsverfahren, Stand 1991, zukünftige Entwicklungen in Deponiesickerwasserreinigung," in *Tagung Deponiesickerwasserreinigung,* München, April 16, 1991.
281. H. Schwarzer: "Oxidative Abwasser-Reinigung mit Wasserstoffperoxid", *GWF Gas Wasserfach: Wasser/Abwasser* **129** (1988) no. 7, 484 – 491.
282. N. Amsoneit: "Wiedergewinnung und Verwertung produktionsspezifischer Abfälle aus Oberflächenbehandlungsanlagen mit anorganischen Stoffen," *gi Gesundheitsingenieur* **5** (1989) 251 – 255.
283. M. Modell: "Supercritical-Water Oxidation," in H. M. Freeman (ed.): *Standard Handbook of Hazardous Waste Treatment and Disposal,* McGraw-Hill, New York 1989.
284. N. Amsoneit: "Einsatz biologischer Verfahren bei der Behandlung flüssiger Abfälle," *Entsorgungspraxis* **4** (1991) 182 – 187.
285. N. Amsoneit: Minderung von Emissionen organischer Luftschadstoffe bei Abwasserbehandlungsanlagen am Beispiel der neuen Abwasserreinigungsanlage des Zweckverbandes Sondermüll-Entsorgung Mittelfranken (ZVSMM), *Galvanotechnik* **82** (1991) no. 9, 3141 – 3146.
286. Verordnung über Anlagen zum Lagern, Abfüllen und Umschlagen wassergefährdender Stoffe und die Zulassung von Fachbetrieben (Anlagen- und Fachbetriebsverordnung – VAwSF) vom 13. 02. 1984, BayRS 753-1-4-J, GVBl., p. 66.
287. G. Härdtle, K. Marek, B. Bilitewski, K. Krijewski: "Recycling von Kunststoffabfällen (Grundlagen – Technik – Wirtschaftlichkeit)," *Müll Abfall,* **24** (1992) no. 27.
288. G. Menges, W. Michaeli, M. Bittner (eds.): *Recycling von Kunststoffen,* Hanser Verlag, München 1992.
289. Wyandotte Chemical, US 344 1616, 1969 (L. C. Pizzini, J. T. Patton).
290. Ford, US 431 6992, 1982 (J. L. Gerlock, J. Braslaw, H. W. Patzelt).
291. E. Grigat, *Kunststoffe,* **68** (1978) no. 5, 284.
292. G. Niederdellmann, H. Hetzel, A. Bergmann-Franke, E. Grigat: "Hydrolyse von Kunststoffabfällen," Teilprojekt 6 des Forschungsprogrammes *Wiederverwertung von Kunststoffabfällen* (1979).
293. J. Gerlock, J. Braslaw, M. Zinbo, *Ind. Eng. Chem. Process Des. Dev.* **23** (1984) 545.
294. J. Braslaw, J. Gerlock, *Ind. Eng. Chem. Process Des. Dev.* **23** (1984) 552.
295. Firma REGRA-Recycling, personal communication, Jan. 1994.
296. R. J. Ehring: *Plastics Recycling, Products and Processes,* Hanser Verlag, München 1992.
297. Du Pont US 407 8143, 1978 (A. I. Molik, E. E. Most).
298. Michigan Technological University, US 454 2239, 1985 (R. A. Lamparter, B. A. Barnu, D. R. Johnsrud).
299. Toray Industries US 460 9680, 1986 (A. A. Fujita, S. M. Sato, M. M. Murakami).
300. COCA-COLA, publication, Essen, Feb. 1994.
301. Celanese Mexicana, US 460 5762, 1986 (W. J. Mandoki).
302. W. F. Müller, Kunststoffberater; **38** (4) (1993); 57.
303. Frau Dr. Seliger, BASF, personal communication, March 3, 1994.
304. G. Bauer: "Verfahren zur Umwandlung polymerer Mischabfälle aus der Autositz-Produktion in Polyole," *FAT-Schriftenr.* **86,** (1990), May.
305. K. V. Paczkowski: "Kessel als Bestandteil einer Müllverbrennungsanlage – Entwicklung und Entwurf," in K. J. Thomé-Kozmiensky (ed.): *Thermische Behandlung von Haushaltsabfällen,* Reihe: Abfallwirtschaft an der TU Berlin, vol. 2, 1978, pp. 371 ff.
306. H.-G. Kitzerow: "Wirbelschichtfeuerungsanlagen für industrielle Reststoffe," in K. J. Thomé-Kozmiensky (ed.): *Reaktoren der thermischen Abfallbehandlung,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1993, pp. 135 – 164.
307. H.-D. Kuchta: "Feuerfestzustellung von Drehrohrofenanlagen zur Sonderabfallverbrennung," in K. J.

Thomé-Kozmiensky (ed.): *Sonderabfallwirtschaft,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1993, pp. 439 – 454.
308. K. J. Thomé-Kozmiensky: *Thermische Abfallbehandlung,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994, pp. 125 – 281.
309. K. J. Thomé-Kozmiensky, U. Pahl: *Verfahrensvergleich: Rostfeuerung, Wirbelschichtfeuerung, Schwel-Brenn-Verfahren, Thermoselect, Noell-Konversionsverfahren,* Studie im Auftrag der BSR, Thomé-Kozmiensky & Partner, Neuruppin 1994 (unpublished).
310. L. Barniske: "Entwicklungsstand der thermischen Abfallbehandlung in der BRD," in: *Zukunftsaspekte der thermischen Abfallbehandlung im Rahmen abfallwirtschaftlicher Planung,* 16. AWS der Technischen Universität, Berlin 1983.
311. R. Berghoff: "Zur Pyrolyse häuslicher Abfälle," Ph. D. Thesis, Technische Hochschule Aachen 1981.
312. S. Lenz: "Optimale Abfallnutzung durch neuartige Pyrolyseverfahren," *Umwelt* (1979) no. 4, 291 – 292.
313. E. Thomanetz: "Möglichkeiten und Grenzen der Erzeugung von Energie und Rohstoffen durch die Pyrolyse von Abfällen," in: *Zukunftsaspekte der thermischen Abfallbehandlung im Rahmen abfallwirtschaftlicher Planung,* 16. AWS der Technischen Universität, Berlin 1983.
314. K. Sattler (ed.): *Umweltschutz, Entsorgungstechnik,* 1st ed., Vogel-Verlag, Würzburg 1982.
315. G. Tchobanoglons, H. Theisen, R. Ellassen: *Solid Wastes,* McGraw-Hill, New York 1977.
316. C. Nels: "Beitrag der Hausmüllpyrolyse zur flä-chendeckenden Abfallentsorgung," in: *Zukunftsaspekte der thermischen Abfallbehandlung im Rahmen abfallwirtschaftlicher Planung,* 16. AWS der Technischen Universität Berlin 1983.
317. J. Heil: "Einfluß der Pyrolysebedingungen und der Abfallarten auf die entstehenden Produkte bei der Entgasung von Haushaltsabfällen," Dissertation, Technische Universität Berlin 1980.
318. E. Douglas, M. Webb, G. R. Daborn: *The Pyrolysis of Waste and Product Assessment,* Warren Spring Laboratory, Stevenage.
319. B. Bilitewski: "Gezielte Herstellung von Adsorptionskoksen aus Abfällen," Ph. D. Thesis, Technische Universität Berlin 1980.
320. K. J. Thomé-Kozmiensky: *Thermische Abfallbehandlung,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994.
321. G. Kiss, W. Marfiewicz, J. Riegel, R. Stahlberg: "Thermoselect – Energie- und Rohstoffgewinnung aus Entsorgungsgütern," in F. J. Schweitzer (ed.): *Thermoselect-Verfahren zur Ent- und Vergasung von Abfällen,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994, pp. 21 – 56.
322. H.-G. Kitzerow, H.-D. Rompel: "Anwendung der rotierenden Wirbelschicht für die Verbrennung von Haushaltsabfällen," in K. J. Thomé-Kozmiensky (ed.): *Müllverbrennung und Umwelt 4,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1990, pp. 145 –161.
323. M. Schingnitz: "Darstellung des Noell-Konversionsverfahrens," in J. Carl, P. Fritz (eds.): *NOELL-Konversionsverfahren zur Verwertung und Entsorgung von Abfällen,* EF-Verlag für Energie- und Umwelttechnik, Berlin 1994, pp. 25 – 53.

General References

324. B. Bilitewski, G. Härdtle, K. Marek: *Abfallwirtschaft,* Springer Verlag, Berlin 1990.
325. K. Cord-Landwehr: *Einführung in die Abfallwirtschaft,* B. G. Teubner, Stuttgart 1994.
326. K. J. Thomé-Kozmiensky (ed.): *Deponie-Ablagerung von Abfällen,* EF-Verlag, Berlin 1987.
327. K. J. Thomé-Kozmiensky (ed.): *Deponie-Ablagerung von Abfällen 2,* EF-Verlag, Berlin 1988.
328. K. J. Thomé-Kozmiensky (ed.): *Deponie-Ablagerung von Abfällen 3,* EF-Verlag, Berlin 1989.
329. K. J. Thomé-Kozmiensky (ed.): *Deponie-Ablagerung von Abfällen 4,* EF-Verlag, Berlin 1990.
330. P. Bacchini (ed.): *The Landfill,* Springer Verlag, Berlin 1988.

Specific References

331. Landesamt für Wasser und Abfall NRW: Entwurf einer Richtlinie über die Untersuchung und Beurteilung von Abfällen, Düsseldorf 1987.
332. Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit: Technische Anleitung zur Verwertung, Behandlung und sonstigen

Entsorgung von Siedlungsabfällen (TA Siedlungsabfall), Bonn 1993.
333. Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit: Technische Anleitung zur Ablagerung, chemisch-physikalischen und biologischen Behandlung, Verbrennung und Ablagerung von besonders überwachungsbedürftigen Abfällen (TA Abfall), Bonn 1991.
334. K. Stief: "Das Multibarrierenkonzept als Grundlage von Planung, Bau, Betrieb und Nachsorge von Deponien," *Müll Abfall* **18** (1986) no. 1, 15.
335. K. Stief: "Multi-Barrier Concept in West Germany," in T. H. Christiansen, R. Cossu, R. Stegmann (eds.): *Sanitary Landfilling Process, Technology and Impact '89,* Academic Press, 1989, p. 549.
336. ATV-Arbeitsgruppe 7.2.26: "Sickerwässer aus Industrie- und Sonderabfalldeponien," *Korresp. Abwasser* **33** (1986) no. 9, 829.
337. H.-J. Ehrig: "Beitrag zum quantitativen und qualitativen Wasserhaushalt von Mülldeponien," Veröffentlichungen des Institutes für Stadtbauwesen, TU Braunschweig, H. 26 (1980).
338. H.-J. Ehrig: "Auswirkungen des Deponiebetriebes auf Sickerwasserbelastungen – Messungen an Deponien und Lysimetern," Veröffentlichungen des Institutes für Stadtbauwesen, TU Braunschweig, H. 39 (1985).
339. H.-J. Ehrig: "Wasserhaushalt von Deponien," in [326], p. 311.
340. H.-J. Ehrig: "Qualität und Quantität von Deponiesickerwasser," *Entsorgungs-Praxis Spezial* **2** (1988) 100.
341. H.-J. Ehrig: "Sickerwasser aus Hausmülldeponien – Menge und Zusammensetzung," in G. Hösel, W. Schenkel, H. Schnurer (eds.): *Müllhandbuch,* Erich-Schmidt Verlag, Berlin 1989, KZ 4587.
342. H.-J. Ehrig: "Wasserhaushalt und Langzeitemissionen von Deponien," in K. Wiemer (ed.): *Abfallwirtschaft 5,* Universität Kassel 1990.
343. H.-J. Ehrig: "Water and Element Balances of Landfills," in [330], p. 83.
344. H.-J. Ehrig, T. Scheelhaase: "Auswirkungen der veränderten Abfallzusammensetzung auf die Qualität des Sickerwassers," *8. ZAF-Seminar Sickerwasser aus Mülldeponien,* TU Braunschweig 1993.
345. L. M. Johannessen, personal communication (1994).
346. O. Hjelmar, personal communication (1994).
347. K. Knox: "A Review of Technical Aspects of Co-disposal," Report prepared for the UK Department of the Environment, 1989.
348. K. Knox, personal communication (1994).
349. H. Robinson: "A Review of Landfill Leachate Composition in the UK," Department of the Environment, 1994, in press.
350. K. Westlake, M. Sayce, T. Fawcett: "A Study of the Type and Scale of Environmental Impacts from Landfills Accepting Predominantly Wastes Other than Domestic," Contact Report for the UK Department of the Environment, Report AEA-EE-0240, Harwell Laboratory, 1991.
351. H.-J. Ehrig, T. Scheelhaase: "Überblick über Stand und Entwicklung der Behandlung von Sickerwässern aus Abfallablagerungen," *4. Symposium Industrieabwässer,* Friedrichrhoda 1993.
352. H.-J. Ehrig: "Untersuchungen zur Gasproduktion aus Hausmüll," *Müll Abfall* **18** (1986) no. 5, 173.
353. H.-J. Ehrig: "Gasprognose bei Restmülldeponien in Rettenberger," in R. Stegmann (ed.): *Trierer Berichte zur Abfallwirtschaft,* Economica Verlag, Bonn 1991, no. 2.
354. H.-J. Ehrig: "Prediction of Gas Production from Laboratory Scale Tests," *Proceedings of the 3rd International Landfill Symposium,* Sardinia 1991.
355. G. Rettenberger: "Ziele und Verfahrenstechnik der Deponieentgasung," in G. Rettenberger (ed.): *Deponiegastechnik,* vol. 4, Economica Verlag, Bonn, 1992, p. 7.
356. Bundesgesetz vom 13. Aug. 1980 und Einigungsvertraggesetz vom 23. Sept. 1990.
357. Gesetz über die Verminderung und Entsorgung von Abfällen (Abfallgesetz AbfG) vom 27. Aug. 1986.
358. Gesetz über die friedliche Verwendung der Kernenergie und den Schutz gegen ihre Gefahren (Atomgesetz) in der Fassung der Bekanntmachung vom 15. Juli 1985.
359. Zweite allgemeine Verwaltungsvorschrift zum Abfallgesetz (TA Abfall) part 1: Technische Anleitung zur Lagerung, chemisch/physikalischen und biologischen Behandlung, Verbrennung und Ablagerung von besonders überwachungsbedürftigen Abfällen vom 10. April 1990, geändert durch AVV vom 17. Dec. 1990.

360. Anforderungen an die stoffliche Verwertung von mineralischen Reststoffen/Abfällen als Versatz unter Tage; Technische Regeln für deren Einsatz von bergbaufremden Reststoffen/Abfällen als Versatz;Entwurf des Ad-hoc-Arbeitskreises Bergbauliche Hohlräume und Abfallentsorgung vom 11. Oct. 1994.
361. G. Fürer: "Entsorgung von Fremdabfällen in bergbaulichen Rahmenbedingungen," *Erzmetall* **42** (1989) no. 5, 189 – 196.

Waste Gas Desulfurization → Sulfuric Acid and Sulfur Trioxide

High-Throughput Screening

Rodney Turner, EVOTEC BioSystems GmbH, Hamburg, Federal Republic of Germany
Sylvia Sterrer, EVOTEC BioSystems GmbH, Hamburg, Federal Republic of Germany
Karl-Heinz Wiesmüller, EVOTEC BioSystems GmbH, Hamburg, Federal Republic of Germany

1. Introduction to Drug Discovery and Screening 1
1.1. A Rationale for new Methods in Drug Discovery 1
1.2. Sources of Compounds for High-Throughput Screening 2
1.3. Demands on High-Throughput Screening 2
2. Compounds 3
2.1. Novel Approaches to Generate Chemical Diversity 3
2.2. Microreactor Systems 3
2.3. Limitations of Pooled Compound Screening Strategies 5
3. Screening Technology 5
3.1. Chemistry, Biology, and Technology . 5
3.2. Primary, Secondary, and Tertiary Screening 6
3.3. Miniaturization, Automation and Homogeneous Assays 7
3.4. Detection Technologies for High-Throughput Screening 8
3.5. Examples of Specific Fluorescence Detection Technologies for High-Throughput Screening 9
3.6. Case Study EVOscreen 10
4. References 12

List of Abbreviations

HTS	high-throughput screening
uHTS	ultra-high-throughput screening
FCS	fluorescence correlation spectroscopy
FRET	fluorescence resonance energy transfer
FP	fluorescence polarization
M2N	micro-to-nano
SAR	structure activity relationship
ADME	absorption, distribution, metabolism, excretion
PK	pharmacokinetics
CV	coefficient of variation

1. Introduction to Drug Discovery and Screening

1.1. A Rationale for new Methods in Drug Discovery

Combinatorial chemistry (→ Combinatorial Chemistry) has been one of the most important keywords in the drug discovery field since the beginning of the 1990s [1–3]. Together with genomics and proteomics [4, 5], the other driving force of the current biotechnology boom, combinatorial chemistry offers the promise of new and better medications. These two new driving forces promise to lead to an explosive increase in both the ability to identify new therapeutic targets (largely thanks to genomics), and the ability to provide tailor-made molecules which can rapidly enter the drug development pipeline (the promise of combinatorial chemistry). The simultaneous, dramatic increase in both the number of targets and potential hit compounds is the reason for what has been called a change in the drug discovery paradigm [6]. To fully capitalize on the promise of both genomics and combinatorial chemistry, the drug discovery process must change in a fundamental way.

Rather than screening 50 chemical compounds against a particular target, combinatorial approaches often require that thousands of compounds be screened. And because of the preponderance of targets, that screening must take place in days rather than months. Out of necessity, high-throughput screening will be a key factor in realizing the potential of new drug discovery strategies by building a bridge between the increased number of targets and the vast number of compounds to be screened.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b13_b01

Not only does screening throughput need to be increased, but novel assay techniques and detection technologies must be employed in order to provide more information about the suitability of specific compounds earlier in the screening process. This combination of quality and quantity must be achieved in order to fully take advantage of the increased potential for target identification offered by genomics and proteomics and the potential for direct, rapid access to novel chemical compounds offered by combinatorial chemistry and combinatorial biology [7].

By employing new methods for the synthesis of discrete molecules, modern combinatorial chemistry has surpassed the capacity of traditional screening strategies in the search for new active substances. Whether screening libraries of high diversity, such as combinatorial peptide and organic compound libraries, or large libraries of individual natural or rationally designed compounds, modern screening strategies dictate that $10^5 - 10^6$ compounds be screened with a single assay system. Using 96-well microtiter plate automated screening technology, a month would be required to screen 10^5compounds at a rate of 50 plates per day. This rate must be cut by roughly a factor of ten for a laboratory to remain competitive in today's pharmaceutical research environment.

1.2. Sources of Compounds for High-Throughput Screening

Natural product libraries, shelf compounds, libraries from chemically modified natural products and combinatorial compound collections, including peptides, peptidomimetics and small organic molecules, are well established as sources for new lead structures (for a definition of the term lead, see → Combinatorial Chemistry, Chap. 1.). Solution-phase or solid phase synthesis with a variety of chemical reactions and scaffolds allows libraries to be directed at almost any class of targets. Combinatorial chemistry is used in structure-based drug design, for lead optimization and to generate highly diverse compound collections for random screening. Methods for their reproducible, effective and fully automated synthesis and analysis have been established and are dependent on the heterogeneity of the mixtures [8]. As the number of products accessible from a given set of components increases exponentially, synthesis strategies had to be developed that allow the highly parallel and simultaneous production of compounds which were generated and tested as mixtures or as individuals [9]. These strategies are supported by new methodologies for diversity measurement and compound selection [10]. Combinatorial and high throughput synthesis have boosted the generation of compounds in many pharmaceutical companies ten- to fifty-fold.

1.3. Demands on High-Throughput Screening

The diversity of compounds accessible by application of combinatorial chemistry places significant demands on screening strategies. An ideal screening system must provide sufficient flexibility to allow a very large number of individual compounds – at least 50 000 compounds per day – be tested against a variety of targets. The goal must also include the wringing of as much information as possible out of a single library synthesis. Due to the desire to use a library a number of times to maximize the use of each synthesized compound, and synthesis strategies which restrict the amount of compound synthesized per round, it is also necessary that high-throughput screening (HTS) systems are able to operate with low amounts of compound. In order to conserve synthetic reagents and the compounds themselves, and achieve biologically relevant concentrations, modern HTS systems must be able to run several assays with picomole quantities of compound in volumes of less than a microliter: Miniaturization is a prerequisite of an efficient HTS system.

Libraries used for lead finding are by nature diverse, but screening places the added demand that this diversity should be describable: A variety of characteristics associated with each individual compound must be registered and stored, chemical and biological information has to be collated. Not only is it necessary to keep track of each compound, the results of screening must also be tracked for a variety of assays. In the future, efficient integrated synthesis and screening systems will be able to take the stream of

compounds directly from synthesis to screening without the need for interim storage. Hits will be resynthesized in scaled-up quantities for downstream studies based on screening and synthesis information alone.

The assays themselves present a significant demand on HTS systems. Established biochemical assays must be readily adapted to miniaturized format. Additionally, assay throughput can be increased by multiplexing strategies: the rapid sequential or simultaneous detection of multiple parameters. This multiplies the amount of information which can be obtained from a single assay and thereby the information that can contribute to the characterization of a single compound. This also means that the lines between primary and secondary screening will become increasingly blurred (see Section 3.2). A variety of read-out modes must be available in any HTS system, not only to enable assay multiplexing and thereby the rapid and efficient characterization of compounds, but also to allow for flexibility in assay design. This flexibility also contributes to the ease with which specific assays can be translated from standard manual formats to miniaturized automated formats.

EVOTEC's goal is to meet all requirements of HTS through the use of EVOscreen (see Section 3.6) to achieve a degree of sensitivity and specificity through high spatial resolution in miniaturized formats, a multiplicity of read out modes all operating in high-throughput mode, with an integrated database system capable of correlating the data required to obtain useful information.

The challenges of high-throughput screening may be summarized as follows:

- Short measurement times (on the order of 1 s)
- Homogeneous assays: "add & read" format
- Improved speed and ease of assay development
- Use of multi-functional assays
- Early elimination of nonpromising hits
- Biological qualification of early leads
- Real-time data handling and evaluation
- Coefficient of variation (CV) of less than 10 %
- Efficient liquid handling on a nanoliter scale
- Integration of compound storage and testing
- High degree of automation and parallelization
- Modular, integrated system architecture

2. Compounds

2.1. Novel Approaches to Generate Chemical Diversity

The construction of a fully automated laboratory where all manual steps – from chemical synthesis to product characterization – are managed by robots is no longer a futuristic dream. Currently, roughly 20 companies are specialized in the design of chemistry robots of different sizes and levels of automation [11, 12]. The robotized laboratory for chemical synthesis is supported by databases for building blocks, for reaction information and for library information.

Compound mixtures and individual compound collections can be designed for screening either in solution [13], or with compounds immobilized on solid supports, such as polypropylene pins [14], or resin beads [15]. Pharmaceutical companies have started to establish a database for published and proprietary procedures for solid phase organic synthesis which can be downloaded to their automated laboratory for chemical synthesis.

2.2. Microreactor Systems

Today, most combinatorial and high throughput synthesis strategies are calculated for 10 mg, 10 – 20 μM, yield. In the first generation of microreactors, single compound beads or spots on functionalized surfaces have been introduced for the synthesis of pico- to nanomolar amounts of product [16]. As they come into wide use, microreactors will bring about a fundamental change in combinatorial synthesis and screening strategies.

Microreactors are formatted for through-connection to microscale purification units and highly sensitive analytical devices [17] as well as ultra-high-throughput screening (uHTS systems). Nanotiter plates are already being used for direct collection from microfluidic separation units, transferring samples directly to biological assay units. Such integrated labs-on-a-chip – combining microreactors, microscale purification and separation units, and highly sensitive analytical devices – are likely to make

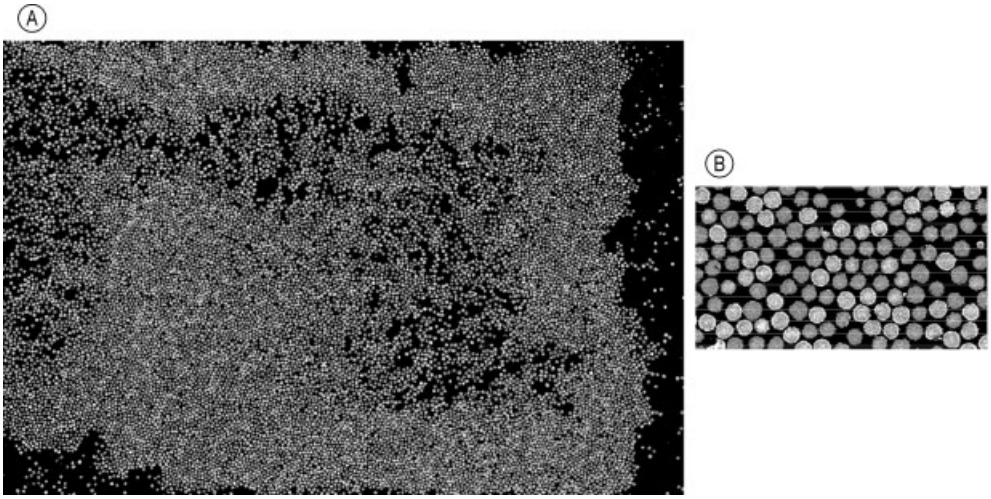

Figure 1. A) Scan of more than 12 000 TentaGel beads using EVOTEC's Bead Scanner/Picker module in the presence of 1 μM fluorescent protein conjugate. B) Close-up of a region of the scan showing the bright fluorescent ring on the positive beads.

compound storage systems obsolete. Test compounds will move directly from synthesis to screening. Furthermore, high-throughput miniaturized synthesis will be designed for parallel synthesis, as well as for the initial scale-up process. To fulfill all of these requirements, all components of integrated microreactors – mixers, valves, and process-control sensors – must be stable to aggressive, highly reactive chemicals.

The benefits of microreactors are many: considerably improved temperature exchange for gas phase reactions; the possibility to carry out the reaction process in an electric field; and simplification of the performance of complex catalytic reactions [18]. These benefits will also provide access to novel chemical structures. Additionally, the bottleneck caused by limited availability of unusual building blocks and scaffolds will be eased or altogether eliminated by microreactor operations and miniaturized screening formats. The advent of microreactor systems for chemical synthesis not only makes direct integration of synthesis and screening possible, but also facilitates the synthesis of novel compounds.

For screening, compounds are removed from the solid support at the point of attachment by cleaving a chemical linker. A variety of chemical linkers have been described, all with the common trait that they enable release of the bead surface compounds under mild conditions (→ Combinatorial Chemistry, Chap. 3.2.1.1.) For identification of specific compounds, several methods for on- or off-bead analysis are available: sequencing of peptides on single beads [16], mass spectrometry of the compounds from one isolated bead [19], chemical tags for decoding by gas chromatography [20], and new tagging strategies, such as DNA tags [21] (→ Combinatorial Chemistry, Chap. 4.2.1.3.) or radiotags (→ Combinatorial Chemistry, Chap. 3.1.1.2.) [22].

Recently, EVOTEC has developed a system for the homogeneous analysis of bead-surface interactions (Fig. 1). This system will enable compound screening to take place without the need to remove compounds from the surface of the bead and thus rapid, on-bead analysis of thousands of beads. Direct recovery of single beads for further analysis is also possible. It eliminates a critical bottleneck, namely the need for deconvolution of solid-phase diversity.

2.3. Limitations of Pooled Compound Screening Strategies

Bioassays are influenced by protein-reactive compounds [23] and tolerate only limited numbers of non-related molecules. False positive and false negative results from screening compound

mixtures have been reported. For example, octa- and hexapeptide mixtures significantly depress binding of the hormone neuropeptide Y to its receptor when measured in a standard competition assay [24].

Many pharmaceutical companies own sample collections of up to a million individual natural or synthetic substances. Combinatorial libraries will enlarge the collections to several million compounds. This explosion in the number of samples and the associated assay costs have forced most screening units to resort to a pooling strategy whereby sometimes thousands of samples are mixed in a single well for primary screening in traditional 96-well formats. This pooling strategy makes the results of screening difficult to interpret, but it is currently the best way to meet necessary throughput levels with conventional assay formats. Different pooling strategies have been experimentally compared using a library of 810 phosphate diester-linked peptide trimers. The results showed that random pooling, positional scanning, and iterative methods identified the same active compound [25]. But applying an artificial procedure where the best compound was hidden in a pool of least active compounds failed to identify the optimal trimer structure [25]. This is the primary reason behind the drive to miniaturization and increased throughput. The screening systems of the future must not only be faster, they must provide more reliable data and be more economical.

3. Screening Technology

3.1. Chemistry, Biology, and Technology

At the screening rates enabled by high and ultra-high throughput applications, reagent consumption is a critical factor. By reducing the volume of reagent required for synthesis and the amount of compound required for analysis, miniaturization during both synthesis and screening phases can lead to significant savings over current methods [26]. Additionally, given the number of compounds to be screened, often of similar chemical structure, novel assay techniques and detection technologies must be employed in order to provide better characterization of specific compounds, and this earlier in the screening process. A collection of synthetic compounds is likely to be used in several different assays – another reason to keep reagent consumption to a minimum. Multiplexing strategies, whereby multiple parameters can be measured by employing multiple assays and/or read-outs in a single sample, will also improve screening efficiency. This combination of quality and quantity must be achieved in order to fully take advantage of the increased potential for target identification offered by genomics and proteomics and the potential for direct, rapid access to novel chemical compounds offered by combinatorial chemistry and combinatorial biology [26]. High- and ultra-high-throughput screening will not only allow realization of the potential of these new technologies, it will also enable better usage of natural compound libraries, long a primary backbone of drug discovery. Screening, therefore, is the link between biology (targets) and chemistry (drug-like molecules). And in the age of genomics and combinatorial chemistry, that link must achieve high throughput. Figure 2 (see next page) illustrates the unification of targets, compounds and assays in screening applications for drug discovery.

The effective use of high-throughput screening for combinatorial applications depends on effective synergy between the traditional disciplines of drug discovery, namely biology, biochemistry, and chemistry, and the technologies associated with combinatorial chemistry [27] and such diverse disciplines as information technology, robotics, physics, and liquid handling. Screening was once the exclusive province of large pharmaceutical companies. Today, screening is increasingly carried out by alliances between biotech ventures, devoted to developing screening strategies, and research laboratories, whether they be large pharmaceutical firms or other biotech ventures, who provide a specific target for screening.

3.2. Primary, Secondary, and Tertiary Screening

Drug screening has traditionally been divided into three phases: primary, secondary and tertiary. In the primary phase, a biological target (often referred to in the pharmaceutical industry as a therapeutic target) is screened against

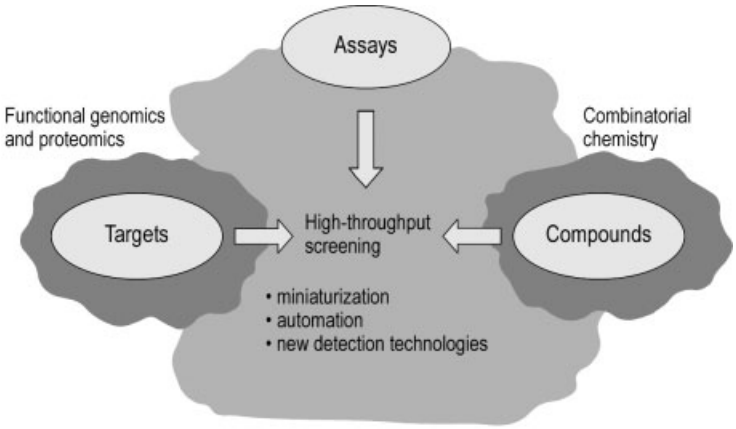

Figure 2. The integration of the three primary components of drug discovery screening: compounds, targets, and assays.

a large number of compounds in a biological assay with the aim of finding hundreds or thousands of compounds which interact (usually in the form of binding) with the target [28]. Compounds selected in the primary screen, usually designated "hits" for a particular target, then enter the secondary phase of screening. In the secondary phase of screening, each hit is characterized as to a range of biological parameters, such as solubility, toxicity, etc. The secondary phase of screening is also where the process of hit optimization occurs, often using focused or directed libraries of structurally-related chemical compounds. The compounds that emerge from the secondary round of screening are then labeled "leads" to designate their status as potential drugs. In the tertiary rounds of screening, compounds are screened for a variety of parameters important to use as a medication: ease of uptake, speed and breadth of distribution throughout the body, modes of metabolism and excretion, and pharmacokinetics. Each phase of this screening process lasts several years in current practice with the entire drug discovery process typically spanning eight to twelve years.

As a result of the dramatic increase in targets and compounds to be screened, traditional screening is currently undergoing a reevaluation. In addition to the move to miniaturization and increases in throughput, increased efficiency is also called for. One way to increase efficiency is to obtain more information at an earlier stage of drug screening.

Primary screening is largely the establishment of a structure-activity relationship (SAR). At this stage of screening, several strategies are typically employed to manage the biological testing of large chemical compound sets against multiple biological targets. Bioassays typically take the form of a competitive binding reaction [28]. For this purpose, the evaluation of a single compound per bioassay per well provides the most unambiguous information and this is precisely why throughput must be increased. The advantages are clear: with a single compound per well, no deconvolution is required and minimal potential for "masking" of bioactivity exists. With one compound per well, the primary bioassay provides extensive structure activity relationship data and additional information for subsequent lead optimization activities. Miniaturization and automation are required to reduce the cost of this approach compared to the alternative strategy of "compound pooling". While the pooling of 3 – 10 compounds per bioassay has been utilized to quickly and efficiently assay large sets of compounds, the primary disadvantage is the need for subsequent deconvolution of positive read-outs, and the potential for masking of one compound's activity by others. By maintaining the one-compound-per-well configuration, the maximum amount of information can be obtained per compound.

Secondary Screening. In secondary screening, multifunctional testing for parameters such as selectivity, toxicity, dose-response is con-

ducted. By combining miniaturization, automation, and multiple detection modes, this process can be moved forward in the drug screening process. New read-out technologies will enable several parameters to be measured simultaneously meaning that assays, rather than compounds, will be "pooled" in a single screening run. Just as conventional strategies led to the pooling of compounds to increase throughput, new detection technologies enable far more information to be obtained in a single measurement, leading to an increase in information and a shortening of screening time. This will make lead evaluation feasible in early stages of the drug discovery process.

Tertiary Screening. Novel types of assay can also be designed for tertiary screening: the profiling of compounds to obtain valuable biological information, such as bioavailability and cellular uptake. To predict whether a molecule has the potential for in vivo pharmacological activity, it is necessary to determine whether the molecule persists in the body long enough at concentrations sufficient to exert the intended pharmacological effect. Such evaluations hinge on assessments of absorption, distribution, metabolism, excretion and pharmacokinetics (ADME/PK). If these evaluations are made in the early stages of drug discovery, resources can be appropriately channeled to those molecules that are likely to succeed in terms of in vivo efficacy. Moreover, where a series of lead compounds is identified by in vitro pharmacological testing, AMDE/PK provides a means of rank-ordering the compounds based on characteristics essential to an effective medication. Such tertiary screens are typically carried out in test animals or in living cells. New detection technologies are being used in functional cellular assays to explore signal transduction, transcriptional control, and even cellular viability. The availability of human liver tissues, cell culture systems (e.g., Caco-2) as a model of the intestinal mucosa, recombinant drug metabolizing enzymes, and cytochrome P450 isoform-selective substrates and inhibitors have provided the tools to effectively conduct such in vitro studies.

Blurring the Lines Between Primary, Secondary, and Tertiary Screening. Because of the difficulty of working with living cells, as well as the difficulty in automating functional assays, cellular assays have been relegated to smaller screens much later in the discovery process. The need to increase the screening speed and the type of information derived means that assays using living cells are now even being used in primary screens. Most assay strategies for use in living cells rely on genetically manipulated (recombinant) cells. However, such over-expressing recombinant cell lines frequently have altered physiological properties compared to primary or nonrecombinant cells, particularly with regard to regulatory mechanisms of signal transduction. Thus they cannot be considered to be ideal for this purpose. For cellular assays, miniaturized systems are capable of providing a superior solution: the combination of sensitivity and miniaturization enables primary cells to be employed. The restrictions resulting from the availability of primary cells are overcome by the very small number of cells needed for each assay (100–1000 cells per well).

3.3. Miniaturization, Automation and Homogeneous Assays

Miniaturization is an inherent characteristic of most modern high-throughput screening systems, but the detection technology is often largely unchanged from many of the read-out technologies used in conventional assays. To provide the most effective synergy between screening technology and assay development, screening systems must provide detection technology which is matched to the demands of miniaturization, on the one hand, and offers significant flexibility of read-out methods, on the other. To fully exploit the advantages offered by miniaturization and enable the highest throughput possible, modern high-throughput screening systems must possess a high degree of automation. The need to precisely deliver assay components in sub-microliter quantities requires not only high-precision liquid handling systems, but automated control of such systems since the volumes handled can barely be seen with the naked eye, much less manually pipetted.

Homogeneous assay methods have been developed which enable "add & read" strategies to be employed. In such homogeneous systems,

no washing, concentration, or separation steps are required. The read-out technologies employed for such systems are discussed in Section 3.4. All assay components, including the target and the compound to be screened, are added to a single well in which the interaction is directly evaluated. Homogeneous methods have been developed to assay the following biological parameters:

1) Molecular binding events: the study of the binding kinetics of molecular interactions, such as receptor-ligand, protein-protein, protein-nucleic acid, and nucleic acid-nucleic acid
2) Molecular processing: the evaluation of the progression and outcome of enzymatic action in assays designed to measure such properties as aggregation, polymerization, degradation, or enzymatic modification of proteins or nucleic acids
3) Vesicular and cellular events: the study of cellular systems for the evaluation of transcriptional activity, signal transduction, membrane-bound or cytoplasmic reactions (potentially involving intracellular navigation), etc.

3.4. Detection Technologies for High-Throughput Screening

Fluorescence-based detection technologies are the method of choice of most modern high-throughput screening systems. Assays based on the use of radioactive tagging, however, still predominate in the area of diagnostics. This is because of a fundamental difference in the types of assays called for in screening and diagnostics. Diagnostic assays require the use of heterogeneous assays (marked by the inclusion of steps to separate unused ligand) to achieve a wide dynamic range of detection and high sensitivity. In screening, on the other hand, the establishment of the structure-activity relationship is accomplished in competitive binding assays readily adaptable to homogeneous formats [28]. For this reason, the use of fluorescent dyes has effectively replaced radioactivity as the tagging method of choice for high-throughput biological assays [29].

The ability to make full use of the fluorescent molecule and all of its properties should be the benchmark of any fluorescence-based assay technology. While most assay strategies make use of only a single fluorescent property, the broad applicability of a screening system requires that flexibility of assay design be extended to the read-out technology by including a variety of detection modes. Although the use of fluorescence in biological assay systems is frequently restricted to the measurement of fluorescence intensity, the measurement of additional fluorescent properties, such as lifetime, polarization, fluorescence energy transfer and quenching can yield a wealth of information from a single measurement. This ability to collect multiple data points per measurement not only provides an internal control, but also contributes to screening efficiency by enabling rapid multi-parameter evaluation of compound-target interactions.

In addition to the use of fluorescence tagging, confocal microscopic optical systems form the foundation of many modern read-out technologies. Using confocal optics, sub-microliter miniaturization may be accomplished without any loss of signal quality. The highly-focused confocal optics reduce the detection volume to as little as one femtoliter. This means that the detection volume, roughly the size of a bacterial cell, is much smaller than a eukaryotic cell or standard miniaturized assay volumes of one microliter. Whereas other methodologies often suffer from the increased contribution of surface effects and from drastically reduced signal-to-noise ratios when assay volumes are reduced, the confocal optics eliminate such limitations. Furthermore, confocal detection technology enables the use of homogeneous assay methods, i.e. those which eliminate washing or so-called "sandwich" strategies (i.e., increasing the assay signal by "sandwiching" the analyte between multiple binding molecules, such as antibodies, necessitating multiple steps of washing). This is a great advantage in screening in that it allows biological systems to be evaluated in close to their natural in vivo state.

Table 1 lists a variety of different fluorescent parameters used as read-out modes for biological assays and their areas of application. Though this list is by no means exhaustive, it gives an indication of the wide variety of information ac-

cessible through fluorescent measurement methods.

Table 1. Multifunctional analysis by confocal fluorescence detection

Detection value	Analytical information
Translational diffusion	molecular mass, shape, binding events, microenvironmental viscosity, oligomerization
Confocal fluorescence lifetime	conformational changes, binding events, potential, electronic environment
Fluorescence depolarization/correlated rotational diffusion	conformational changes, binding events
Fluorescence resonance energy transfer	proximity
Multicolor assays	cellular navigation, multiparameter testing, multiple binding events
Fluorescence recovery after photobleaching	functional reporter binding, membrane fluidity
Fluorescence cross-correlation	molecular mass independent molecular interactions (protein – protein, protein – nucleic acid, nucleic acid – nucleic acid)
Particle number	enzymatic reactions
Triplet transition probability	electronic environment

3.5. Examples of Specific Fluorescence Detection Technologies for High-Throughput Screening

Fluorescence correlation spectroscopy (FCS) takes advantage of differences in the translational diffusion of large versus small molecules [30, 31]. Each molecule which diffuses through the illuminated confocal focus gives rise to bursts of fluorescent light quanta during the entire course of its journey, with each individual burst being registered. The length of each photon burst corresponds to the time the molecule spends in the confocal focus. The photons emitted in each burst are recorded in a time-resolved manner by a highly sensitive single-photon detection device. This detection method achieves single molecule sensitivity, but the fact that diffusion is a random process requires that the diffusion events for a minimum ensemble of molecules be averaged to achieve statistically reliable information. The detection of diffusion events enables a diffusion coefficient to be determined. This diffusion coefficient serves as a parameter to distinguish between different fluorescent species in solution, for example between free or bound ligands. In screening, the diffusion coefficient can be used to determine such factors as concentration or degree of binding. Hereby, confocal optics eliminate any interference from background signals and allow homogeneous assays to be performed. FCS measurements are conducted in a matter of seconds making the technology ideally-suited for high-throughput application [32]. For example, using FCS, not only can a structure-activity relationship be determined between the compound and target, but also the concentration can be determined in a single measurement. This enables dose-response to be evaluated immediately during screening.

Fluorescence resonance energy transfer (FRET) is gaining increasing acceptance as a nonradioactive assay method for screening [33, 34]. FRET capitalizes on the transfer of energy from one fluorescent molecule to another. The distance between fluorescent molecules must be less than 4 – 5 nm [35], which is approximately the diameter of a protein molecule with a molecular mass of 26 000. A major consideration in choosing an assay based on energy transfer is the change in distance that is induced upon ligand binding or enzyme turnover. For example, FRET can be employed for protease assays using peptide substrates labeled with two fluorescent dyes, whereby energy is transferred from one dye to the other. Upon digestion with a protease, the two fluorescent molecules are brought out of close proximity and the loss of fluorescence intensity is measured. Again, using confocal optics, homogeneous assays can be performed using this read-out technology.

Fluorescence Polarization (FP). Since 96-well FP readers have become commercially available, FP has gained importance for screening. The advantages of FP include the need for only one fluorescent label and the easy adaptability to homogeneous assay schemes [36, 37]. In FP, changes in the polarization of fluorescent molecules are measured. As a molecule rotates, the polarization of an attached fluorescent tag can be determined. When such a molecule binds to another, its rotational movement, and thereby the frequency in its changes in polarization, are restricted. Since the polarization shift taking place upon excitation of the fluorophore

is proportional to the fraction of the total fluorophore in the bound state, receptor – ligand assays may need to be performed in the presence of excess receptor, a situation that is not always practical. Enzymes, primarily proteases, but also kinases and phosphatases can also be assayed using FP. Compared with FRET, FP is more useful when the targets are large, since larger changes in rotational mobility can generally be observed. Many manufacturers of FP readers are developing readers compatible with 384-well plates and even higher density formats – opening the way for applicability of this read-out mode to HTS.

Time-Resolved Fluorescence. The analysis of time-resolved fluorescence in homogeneous assays (HTRF) is rapidly gaining acceptance in HTS [29]. The time-resolved analysis of fluorescent emission in bioassays has been made possible by the use of the specific properties of lanthanide ions, especially europium, terbium, samarium, and dysprosium ions. These ions possess unique properties which make then especially attractive for bioassays: they are hydrophilic and have long excited-state lifetimes which enable them to be easily distinguished from other sources of fluorescence emission in a homogeneous sample [28]. Using these lanthanide ions, in the mid-1990s homogeneous assays have been developed for application to HTS [38]. By labeling one component of a competitive binding bioassay with one of these ions, read-out methods are being developed for a range of assays based on the time resolved analysis of fluorescence [28]. Readers are already under development by several companies for densities of up 864 wells in a microtiter plate format. The expense of the ions used for this method is, however, one factor which must be considered in the move to high-throughput applications.

Fluorescence Imaging. By the development of highly sensitive detection systems, imaging technology has become of use in drug discovery. Although this type of detection is based on the measurement of intensity alone, images can be acquired in quick succession enabling rapid changes in cells to be observed. The development of a commercial high-speed, high sensitivity imaging system [39], enabled limitations in the detection of calcium flux to be overcome (the calcium response is short lived, < 1 s). This imaging system is able to collect parallel image data at speeds up to 10 frames per second, at sensitivities that allow detection of calcium transients in cell populations. High-sensitivity fluorescent imaging plate readers have been developed for cellular assays, particularly assays of membrane potential and intracellular calcium release [40]. The instrument uses a high-intensity laser source with narrow depth-of-field optics which helps to eliminate background signals. The combination of fluidics and imaging allows for high-throughput despite the transience of the response. Although fluorescence imaging is widely used, particularly for cellular assays, the influence fo non-specific adsorption effects and variations in signal due to small differences in filling levels limit its application in solution-based assays.

3.6. Case Study EVOscreen

EVOTEC BioSystems has developed a unique platform integrating highly sensitive detection technology and high-precision liquid handling systems in a modular system which will be capable of handling 100 000 assays per day. The core elements of the modular system are: A high performance confocal fluorescence detection unit usable in either a single channel or multi-channel mode; proprietary signal processing protocols based on FCS, and related single molecule-based confocal fluorescence methodologies enabling read-out of fluorescence polarization, lifetime, and intensity; a miniaturized, automated liquid handling system for nano- to low microliter volumes (including pipetting, dispensing and compound retrieval); a rapid, miniaturized multi-replica compound repository providing the link between traditional single-compound library formats and EVOscreen; a micro-separation device (HPLC) coupled to the detection system; and a scanner and picker device for the analysis of combinatorial libraries and for functional genome analysis.

Liquid Handling. The ability to miniaturize assays to the sub-microliter level is largely dependent on the precision of liquid handling systems. Modern pipetting and dispensing systems enable components to be reproducibly

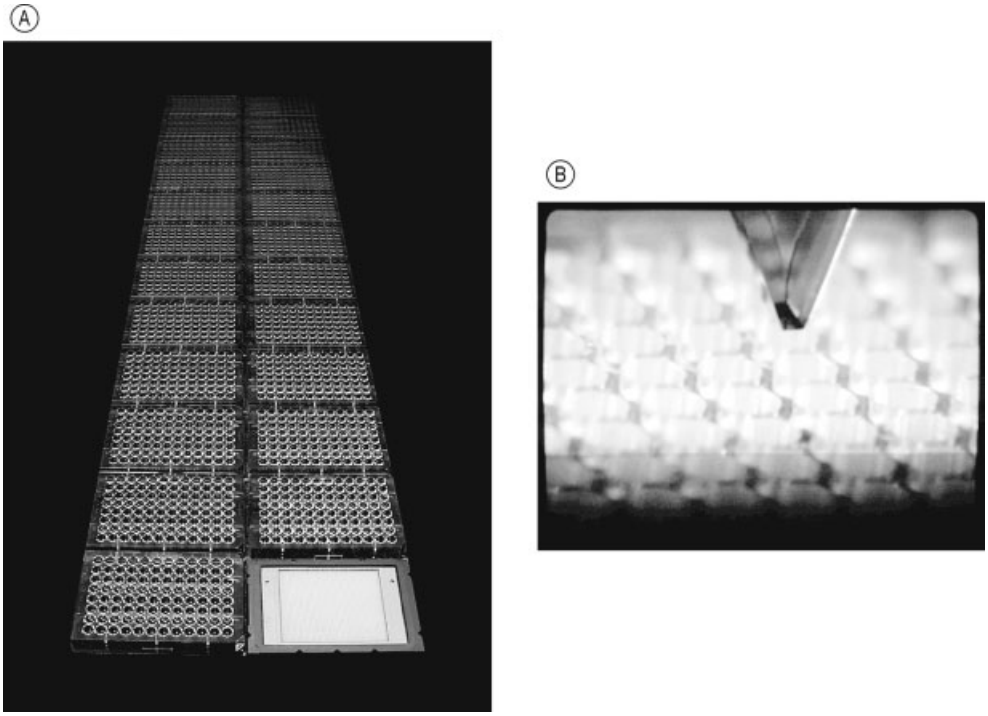

Figure 3. A) Comparison of the reduction in space achieved by using miniaturized sample carriers. 2000 samples may be carried by either the 23 microtiter plates seen in this picture, or a single 2000-well nanocarrier seen in the lower right
B) The head of a nanodispenser over one of the wells of a 2000-well nanocarrier.

added in volumes in the nanoliter range to thousands of samples per plate. The EVOscreen system includes a specially designed micro-to-nano (M2N) unit which is able to reformat samples stored in any conventional format, such as 96-well plates, to the high-density 2000-well nanocarriers (see Fig. 3) used for miniaturized assays. Also included is an automated storage and retrieval system providing a compound repository, as well as transport mechanisms for the many millions of compounds to be used in the screening runs. One specific feature of the EVOscreen liquid handling system is single-drop monitoring and on-line quality assurance.

Acknowledgements. The authors would like to thank Drs. Rolf Günther and Eloisa Lopez-Calle for reviewing the manuscript and for helpful discussions.

4. References

1. M. A. Gallop, R. W. Barrett, W. J. Dower, S. P. A. Fodor, E. M. Gordon, *J. Med. Chem.* **37** (1994) 1233 – 1251.
2. J. S. Früchtel, G. Jung, *Angew. Chem. Int. E. Engl.* **29** (1996) 17 – 42.
3. J. A. Ellmann, *Angew. Chem. Int. E. Engl. Acc. Chem. Res.* **29** (1996) 132 – 143.
4. L. J. Beeley, D. M. Duckworth, *Drug Discovery Today* **1** (1996) 474 – 480.
5. A. Persidis, *Nat. Biotechnol.,* **16** (1998) 393 – 394.
6. A. W. Lloyd, *Drug Discovery Today* **2** (1997) 397 – 398.
7. G. L. Verdine, *Nature* **384** (1996) Supl., 11 – 13.
8. E. M. Gordon, M. A. Gallop, D. V. Patel, *Acc. Chem. Res.* **29** (1996) 144 – 154.
9. D. Madden, V. Krchnak, M. Lebl, *Persp. Drug Discovery Des.* **2** (1995) 269 – 285.
10. P. Willet in A. Czarnik, S. H. DeWitt (eds.): *A Practical Guide to Combinatorial Chemistry,* ACS Books, Washington DC 1997, pp. 17 – 48.

11. R. A. Rivero, M. N. Greco, B. E. Maryanoff in A. Czarnik, S. H. DeWitt (eds.): *A Practical Guide to Combinatorial Chemistry,* ACS Books, Washington DC 1997, pp. 281 – 307.
12. J. F. Cargill, M. Lebl, *Curr. Opin. Chem. Biol.* **1** (1997) 67 – 71.
13. H. An, B. D. Haly, P. D. Cook, *J. Med. Chem.* **41** (1998) 706 – 716.
14. H. M. Geysen, T. J. Mason, *Bioorg. Med. Chem. Lett.* **3** (1993) 397 – 404.
15. J. K. Chen, S. L. Schreiber, *Angew. Chem. Int. Ed. Engl.* **34** (1995) 953 – 969.
16. W. Rapp in G. Jung (ed.): *Combinatorial peptide and nonpeptide libraries – A handbook for the search of lead structures,* VCH Verlagsgesellschaft, Weinheim 1996 pp. 425 – 458.
17. F. Garcia, J. D. Henion, *Anal. Chem.* **64** (1992) 985 – 990.
18. W. Ehrfeld: *Microreaction Technologies, Proceedings of the First International Conference 1998,* Springer-Verlag, Berlin 1998.
19. C. Brummel et al., *Science* **264** (1994) 399 – 402.
20. M. H. J. Ohlmeyer et al., *Proc. Natl. Acad. Sci. USA* **90** (1993) 10922 – 10926.
21. S. Brenner, R. A. Lerner, *Proc. Natl. Acad. Sci. USA* **82** (1992) 5381 – 5383.
22. K. C. Nicolaou et al., *Angew. Chem. Int. E. Engl.* **34** (1995) 2289 – 2292.
23. G. M. Rishton, *Drug Discovery Today* **2** (1997) 382 – 384.
24. K. H. Wiesmüller et al. in P. T. P. Kaumaya, R. S. Hodges (eds.): *Peptides, Proc. 13th Amer. Peptide Symposium,* Escom, Leiden 1996 pp. 998 – 1000.
25. L. Wilson-Lingardo et al., *J. Med. Chem.* **39** (1996) 2730 – 2726.
26. S. Borman, C&EN, **1998** April 6, 47 – 67.
27. S. P. A. Fodor et al., *Science* **251** (1991) 767 – 773.
28. I. Hemmilä, S. Webb, *Drug Discovery Today* **2** (1997) 373 – 381.
29. M. V. Rogers, *Drug Discovery Today* **2** (1997) 156 – 160.
30. M. Eigen, R. Rigler, *Proc. Natl. Acad. Sci. USA* **91** (1994) 5740 – 5747.
31. R. Rigler, *J. Biotechnol* **41** (1995) 177 – 186.
32. M. Auer et al., *Drug Discovery Today* (1998), in press.
33. T. T. Yang et al., *Gene* **173** (1996) 13 – 17.
34. P. R. Selvin, *Methods Enzymol.* **246** (1995) 300 – 334.
35. C. R. Cantor, P. R. Schimmel, *Biophysical Chemistry, San Francisco,* **1980** 451 – 454.
36. W. J. Chekovich, R. E. Bolger, T. Burke, *Nature* **375** (1995) 254 – 256.
37. D. M. Jameson, W. H. Sawyer, *Methods Enzymol.* **246** (1995) 283 – 300.
38. J. M. Kolb, G. Yamanaka, S. P. Manly, *J. Biomol. Screening* **1** (1996) 203 – 210.
39. M. Akong et al., *Soc. Neurosci. Abstr.* **21** (1995) 577.
40. K. Schroeder, B. Neagle, *J. Biomol. Screening* **1** (1996) 75 – 80.

Histamine Antagonists → Antiallergic Agents
Histidine → Amino Acids

Metallocenes

Walter Kaminsky, University of Hamburg, Hamburg, Germany

1. **Introduction** 1
2. **Polyethylene and Polypropylene** 3
3. **Cycloolefin Copolymers** 4
4. **Syndiotactic Polystyrene** 5
5. **Supporting of Metallocene Catalysts** . . 5
6. **References** 6

By application of metallocenes a new catalyst family has been introduced into polyolefin production, the most quickly growing market in the polymer industries. The easy modification of the metallocene catalyst structure brought about improvements in catalyst activities, comonomer incorporation, and polymer properties. It became possible to tailor in a wide range the microstructure of polyolefins such as chain length, molecular mass distribution, polypropylene tacticity, and comonomer distribution.Ethylene/oct-1-ene and cycloolefin copolymers, syndiotactic or isoblock polypropylenes and syndiotactic polystyrene are some types of new industrially produced polymers. The polyolefins synthesized by metallocene catalysts show a narrow molecular mass distribution of 2 and in some cases long chain branches.

1. Introduction

Metallocene – methylaluminoxane (MAO) (→ Aluminum Compounds, Organic) catalysts have opened a frontier in the area of organometallic chemistry, polymer synthesis and processing [1, 2]. A great number of different titanocenes zirconocenes, and hafnocenes have been synthesized that produce tailored polymers of totally different structures and make the control of polymer tacticity, molecular mass and molecular mass distribution more efficient. New kinds of homopolymers, copolymers and elastomers can be produced [3 – 5].

Metallocene catalysts are soluble in hydrocarbons. These catalysts show only one type of active site (→ Polyolefins) and their chemical structure can be easily changed. This allows (1) an accurate prediction of the properties of the resulting polyolefins by knowing the structure of the catalyst used during their manufacture and (2) the control of the resulting molecular mass and mass distribution, comonomer content and tacticity by careful selection of the appropriate reactor conditions. In addition, the activity of these catalysts is ten to hundred times higher than that of the classical Ziegler – Natta systems. Metallocenes, in combination with the conventional aluminumalkyl cocatalysts used in Ziegler systems, are indeed capable of polymerizing ethylene, but at a very low activity. Only with the application of methylaluminoxane was it possible to enhance the activity by a factor of 10 000 and more [6, 7]. Therefore, MAO plays an crucial part in the catalysis with metallocenes.

Methylaluminoxane is a compound in which aluminum and oxygen atoms are arranged alternately and free valences are saturated by methyl substituents. It is produced by careful partial hydrolysis of trimethylaluminum (→ Aluminum Compounds, Organic, Chap. 3.1) and, according to investigations by Sinn [8] and Barron [9], it consists mainly of units of the basic structure $[Al_4O_3(CH_3)_6]$, which contains four aluminum, three oxygen atoms and six methyl groups. As the aluminum atoms in this structure are coordinatively unsaturated, the basic units join together forming clusters and cages. These have molecular masses ranging from 1200 to 1600 and are soluble in hydrocarbons.

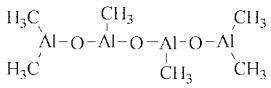

If metallocenes, especially zirconocenes, are treated with MAO, catalysts are obtained that allow the polymerization of up to 100 t ethylene

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b16_b36.pub2

per gram of zirconium. At such high activities (i.e., small quantities of catalyst required) the catalyst can remain in the product. The insertion time (time needed for the insertion of one molecule of ethylene into the growing chain) amounts to some 10^{-5} s only.

It is generally assumed that MAO first undergoes a fast ligand exchange reaction with the metallocene dichloride, thus forming the metallocene methyl and dimethyl compounds. In a second step, either Cl^- or CH_3^- is abstracted from the metallocene compound by an Al-center in MAO, thus forming a metallocene cation and a MAO anion [10, 11]. The alkylated metallocene cation represents the active center. Meanwhile, other weakly coordinating cocatalysts, such as tetra(perfluorophenyl)borate anions $[(C_6F_5)_4B]^-$, have been successfully applied to the activation of metallocenes [12 – 14].

Methylation

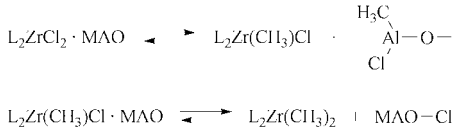

Activation

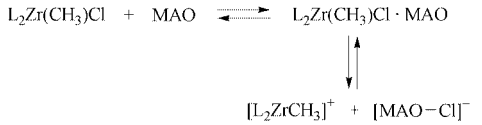

Further milestones were reached in 1982, when Brintzinger [15] synthesized chiral bridged metallocenes and in 1984, when Ewen [16] demonstrated that partially isotactic polypropylene can be produced with appropriate titanocene catalysts. Highly isotactic material was obtained with analogous zirconocenes [17]. After this discovery, an avalanche-like development of industrial and scientific research in the metallocene sector commenced.

Polyolefins, with different microstructures and characteristics, can be custom-made just by varying the ligands of the metallocene (Table 1) [18, 19]. By combining different olefins and cycloolefins with one another, the range of characteristics can be broadened further.

Table 1. Structures of selected metallocenes used for polymerization of olefins

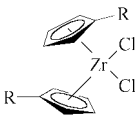

1 R = H
2 R = neomenthyl

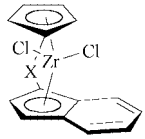

3 X = Me_2C
4 X = Ph_2C

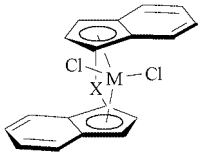

5 M = Zr, X = C_2H_4
6 M = Hf, X = C_2H_4
9 M = Zr, X = Me_2Si

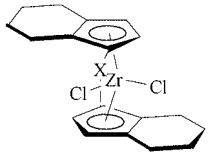

7 = C_2H_4
10 X = Me_2Si

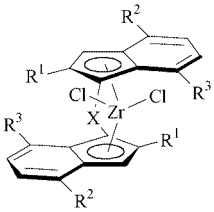

8 X = C_2H_4, $R^1 = R^2 = R^3$ = Me
11 X = Me_2Si, $R^1 = R^2 = R^3$ = Me
12 X = Me_2Si, R^1 = Me, R^2 = Ph, R^3 = H

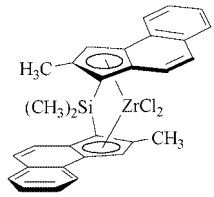

13

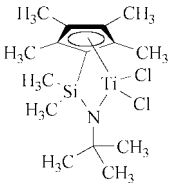

14

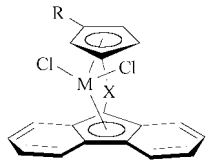

15 M = Zr, X = Me_2C, R = H
16 M = Hf, X = Me_2C, R = H
17 M = Zr, X = Ph_2C, R = H
18 M = Zr, X = Me_2C, R = Bu^t

Using metallocene catalysts, it was possible for the first time to produce polyethylenes, polypropylenes, and copolymers with narrow molecular mass distributions ($M_w/M_n = 2$) [6, 21], syndiotactic polypropylene (in technical scale amounts) [22], syndiotactic polystyrene [23], cyclopolymerizates of 1,5-hexadiene [24], cycloolefin copolymers (COC) with high catalytic activity [25], optically active oligomers [26] and composite materials of biomass, powdered metals, nanoparticles, and polyolefins [27].

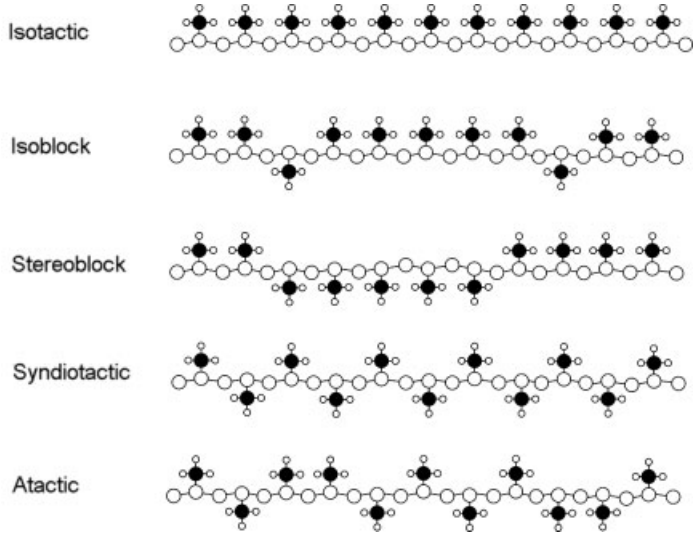

Figure 1. Microstructures of polypropylenes of different tacticities, modified Fischer projection.

2. Polyethylene and Polypropylene

The polymerization of ethylene and propene with selected metallocenes under identical conditions is summarized in Table 2. It is evident that the unsubstituted zirconocene Cp_2ZrCl_2 (**1**) has a high activity in the polymerization of ethylene. Hafnocenes are less active than analogous zirconocenes. Surprisingly, the sterically congested and highly substituted metallocene **11** (see Table 1) displays the highest activity in the polymerization of ethylene among the compounds listed in Table 2 [28]. This leads to the conclusion that electronic effects are particularly important for the insertion reaction. The molecular mass of polyethylene and polypropylene (PP) can be adjusted to lower values by raising the polymerization temperature, by addition of small amounts of hydrogen, or by decreasing the monomer concentration. Apart from isotactic and atactic polypropylene, also syndiotactic and the isoblock and stereoblock materials can be synthesized for the first time in large quantities and high purity (Fig. 1). Syndiotactic polypropylene can be synthesized by complexes containing fluorenyl and cyclopentadienyl ligands bridged by X (compounds **15** – **17**, Table 2) [22]. The block lengths of equal tacticity in isoblock PP (in which the methyl groups extend in one direction) can be varied in a broad range on the nanoscale (4 – 100 units). This leads to microcrystalline PP which is suitable for production of transparent films. The development of applications for elastic stereoblock PP has just begun [29, 30].

By inserting a silyl bridge and substituting the indenyl ligands in zirconocene **1**, the group at Hoechst in Frankfurt and BASF in Ludwigshafen were able to enhance the activity and stereoselectivity of the isotactically working catalyst considerably (compounds **11** – **13**). These catalysts afford isotactic PP of high molecular mass and a melting point of 161 °C [31]. Polypropylenes made by metallocenes exhibit distinct differences to conventionally produced polypropylenes, such as narrow molecular mass distributions, higher stiffness, and greater tensile strength. This is caused not only by the more uniform structure but also by the extremely low fractions of oligomeric products with low molecular mass. These fractions amount to less than 0.1 %, compared to 2 – 4 % in Ziegler – Natta PP. Copolymers of ethylene and oct-1-ene produced by analogous "single site" catalysts **14** of Exxon and Dow in the USA (→ Polyolefins) have already established a market in the packaging industry and in the field of medical applications [32].

Metallocene catalysts will become extremely important in the manufacturing of elastomers with elastic properties similar to rubber. The copolymers of ethylene and propene with a molar ratio of 1 : 0.5 to 1 : 2 are of great industrial interest. These ethylene – propene (EP) polymers

Table 2. Homopolymerization of ethylene and propene at 30 °C, 2.5 bar monomer pressure, 6.25×10^{-6} mol/L metallocene concentration, molar MAO : metallocene ratio = 250 : 1

	Ethylene		Propene		
Catalyst	Activity[a]	Molecular mass PE, kg/mol	Activity[a]	Molecular mass PP, kg/mol	mmmm[b], %
1	60 900	620	140	2	7
2	12 200	1 000	170	3	59
3	1 150	25	180	3	49
5	41 100	140	1 690	323	95
6	2 900	480	610	446	94
7	22 200	1 000	1 220	24	98
8	78 000	190	750	418	> 99
9	36 900	260	1 940	79	97
10	30 200	900	7 700	44	95
11	111 900	250	3 800	192	94
12	16 600	730	15 000	650	99
13	7 600	450	6 100	380	98
15	2 000	500	1 550	159	0.6
16	890	560	130	750	0.7
17	2 890	630	1 980	729	0.4

[a] Measured in kg polyolefin per mol metallocene per h per mol/L monomer.
[b] Isotacticity: mmmm (*meso*) pentads as determined by ^{13}C-NMR spectroscopy.

show elastic properties and, together with 2 – 5 wt % of dienes as third monomers they are used as elastomers (EPDM) [33, 34] (→ Rubber, 3. Synthetic). One of the biggest industrial plant with a capacity of 400 000 t/a for ethylene – propene copolymers was erected by Nova Chemicals, Canada.

The regiospecificity of the metallocene catalysts in regards to propene leads exclusively to the formation of head-to-tail enchainments. 5-Ethylidene-2-norbornene, for example, polymerizes via vinyl polymerization of the cyclic double bond and the tendency to branch is low.

High branching, which is caused by the incorporation of long-chain olefins into the growing polymer chain, is obtained with a new class of silyl bridged amido(cyclopentadienyl)titanium compounds **14** [35]. These catalysts, used by Dow and Exxon in combination with MAO or borates, incorporates oligomers with vinyl endgroups, which are formed during polymerization by β-hydrogen transfer, resulting in long chain branched polyolefins [32].

Long branched copolymers of ethylene with oct-1-ene (polyolefin elastomers, POE) show elastic properties as long as the oct-1-ene content is > 20 % [36]. Other elastomers with different microstructures can be synthesized from dienes [37].

3. Cycloolefin Copolymers

Metallocene catalysts are particularly important for the polymerization of cycloolefins (cyclopentene, norbornene and their substituted compounds). In this process, only the double bond is opened and not the ring. Crystalline polycycloolefins have extremely high melting points of at least 380 °C, being sometimes higher than the decomposition temperature [38 – 40].

While homopolymerization of cyclopentene results in 1,3-enchainment of the monomer units norbornene is inserted in 1,2-enchainment as is usual for polymerization of olefins [41].

The problems of processing that arise from the high melting temperatures of cyclopentene and norbornene homopolymers can be solved by copolymerizing cycloolefins with ethylene, for example. The insertion of norbornene units into the growing polymer chain proceeds very easily. The copolymerization parameter r_1 determines how much faster ethylene is inserted than norbornene if the last monomer unit inserted was ethylene. The value of r_1 is only between 2.0 and 3.4 which shows that it is easy to incorporate this huge monomer. Table 3 compares activities and the incorporation of norbornene for different catalysts.

The metallocene [Me_2C(*tert*-BuCp)(Flu)] $ZrCl_2$ **18** shows not only high activities for the copolymerization of ethylene with norbornene,

Table 3. Copolymerization of norbornene (Nor) and ethylene by different metallocene – MAO catalysts at 30 °C *

Catalyst	Time, min	Activity, kg $mol^{-1}h^{-1}$	Incorporation of Nor, %
Cp_2ZrCl_2	30	1 200	21.4
$[En(Ind)_2]ZrCl_2$	10	9 120	26.1
$[Me_2Si(Ind)_2]ZrCl_2$	15	2 320	28.4
$[En(IndH_4)_2]ZrCl_2$	40	480	28.1
$[Me_2C(Flu)(Cp)]ZrCl_2$	10	7 200	28.9
$[Ph_2C(Flu)(Cp)]ZrCl_2$	10	6 000	27.3
$[Ph_2(Ind)(Cp)]ZrCl_2$	15	2 950	33.3

* Conditions: MAO/Zr = 200, [Zr] = 5×10^{-6} mol/L; ethylene pressure: 2 bar, [Nor] = 0,05 mol/L. Cp = Cyclopentadienyl. En = Ethenyl. Ind = Indenyl. Flu = Fluorenyl.

but also effects an alternating structure in the polymer [42].

It is impossible to produce copolymers with more than 50 mol % of norbornene. The melting point of the alternating copolymer depends on the molar ratio of norbornene units in the polymer while the glass transition temperature is nearly independent on this value. A maximum melting point of 320 °C was achieved.

Cycloolefin copolymers (COC) have an excellent transparency and a very high continuous service temperature. From cycloolefin insertion rates of 10 mol % upwards, these copolymers are no longer crystalline but amorphous. They are very resistant towards solvents and chemicals, they exhibit high softening temperatures (glass temperatures of up to 200 °C) and can be processed like thermoplastics. A further peculiarity of these materials is their extremely low light absorbance, which makes them suitable for optoelectronic applications. Norbornene – ethylene copolymers are most interesting for technical uses because of the easily available monomers. Such COC polymers have already been used for the production of compact discs in Japan and films in Germany. 30 000 t/a of COC materials are produced by Ticona, Germany. Applications for optoelectronic data transfer and storage as well as for other areas of high technology have been provided for.

4. Syndiotactic Polystyrene

Idemitsu was able to demonstrate that especially titanium compounds combined with MAO are capable of polymerizing styrene in a syndiotactical manner [23]. Moreover, cyclopentadienyl titaniumtrichlorides ($CpTiCl_3$) have been proven to be remarkably active catalysts in this case [43]. Syndiotactic polystyrene is crystalline and has a melting point of 275 °C, which nearly makes it a high performance plastic.

If fluorinated complexes are employed such as pentamethylcyclopentadienyl titaniumtrifluoride, the activity can be improved by a factor of 30 [44] compared to chlorinated compounds. At the same time the molecular mass rises from 169 000 to 660 000.

The copolymerization of styrene with ethylene, as examined by MÜLHAUPT, leads to new copolymers with special properties and applications [45]. Syndiotactic polystyrene has already been produced in technical amounts by Idemitsu and Dow.

5. Supporting of Metallocene Catalysts

Only few polymerization processes use metallocene catalysts in a homogeneous form. Supported metallocene catalysts are preferred for the production of polyethylene or isotactic polypropylene on an industrial scale. In order to use them in existing technical processes (drop-in technology) by exchanging the conventional Ziegler – Natta catalysts, the metallocenes have to be applied to a powdery, insoluble substrate. Therefore, the metallocenes have to be absorbed on silica, alumina, magnesium dichloride or other supports. Different methods are possible [46 – 52], for instance: (1) initial absorption of MAO on the support with subsequent addition of metallocenes in a second step, which is most frequently used. These catalysts are used in combination with additional MAO or aluminum alkyls in the polymerization. (2) Another way is either absorption and immobilization of the metallocene or covalent bonding by a spacer to the

support surface as the first step. Then, after addition of MAO, this catalytic system is used in the polymerization process. Both procedures afford different catalysts and these, in turn, produce polyolefins with different properties.

The polymers obtained by method (1) are very similar to those obtained by the homogeneous system. Each metallocene on the support forms an active center and the starting point for the growth of a polymer chain. As the active sites on the surface of each catalyst grain are identical, all chains grow uniformly resulting in polymers with narrow molecular mass distributions. If the metallocene is linked to the support first, different types of absorption occur. In addition, a large part of the metallocene is destroyed by acid centers. The different absorption bonds lead to different active sites. Therefore, the activity is much lower than in the case of the homogeneous system and the molecular mass distribution of the polymer produced is much broader.

In the late 1990s other complexes of early transition metals such as titanium, zirconium, and chromium were discovered for the active polymerization of olefins [53]. In these complexes the central atom is surrounded by salicyl aldimine or imino phenolate ligands. Other new complexes were those of late transition metals — nickel, palladium, cobalt, and iron — with diimine ligands [54]. All these catalyst complexes need MAO or fluorinated borates for activation similar to metallocenes.

6. References

1. W. Kaminsky, *Macromol. Chem. Phys.* **197** (1996) 3907.
2. H. H. Brintzinger et al., *Angew. Chem.* **107** (1995) 1255.
3. J. Scheirs, W. Kaminsky (eds.): *Metallocene-Based Polyolefins*, Vol I + II, Wiley, Chichester 2000.
4. R. Blom et al. (eds.): *Organometallic Catalysts and Olefin Polymerization*, Springer, Berlin 2001.
5. J. R. Severn et al., *Chem. Rev.* **105** (2005) 4073.
6. H. Sinn, W. Kaminsky, *Adv. Organomet. Chem.* **18** (1980) 99.
7. W. Kaminsky, *J. Polym. Sci. A Polym. Chem.* **42** (2004) 3911.
8. H. Sinn, *Macromol. Symp.* **97** (1995) 27.
9. Y. Koide, S. G. Bott, A. R. Barron, *Organometallics* **15** (1996) 2213.
10. J. J. Eisch, S. I. Bombrick, G. X. Zheng, *Organometallics* **12** (1993) 3856.
11. R. F. Jordan, W. E. Dasher, S. F. Echols, *J. Am. Chem. Soc.* **108** (1986) 1718.
12. C. Sishta, R. M. Hathorn, T. J. Marks, *J. Am. Chem. Soc.* **114** (1992) 114.
13. F. Song, M. D. Hannant, R. D. Cannon, M. Bochmann, *Macromol. Symp.* **213** (2004) 173.
14. J. K. G. Erker, R. Fröhlich, *J. Am. Chem. Soc.* **119** (1997) 11165.
15. F. R. W. P. Wild, L. Zsolnai, G. Huttner, H. H. Brintzinger, *J. Organomet. Chem.* **232** (1982) 233.
16. J. A. Ewen, *J. Am. Chem. Soc.* **106** (1984) 6355.
17. W. Kaminsky, K. Külper, H. H. Brintzinger, F. R. W. P. Wild, *Angew. Chem. Int. Ed. Engl.* **24** (1985) 507.
18. W. Spaleck et al., *Organometallics* **13** (1994) 954.
19. W. Kaminsky, A. Laban, *Appl. Catal. A General* **222** (2001) 47.
20. H. G. Alt, J. S. Han, U. Thewalt, *J. Organomet. Chem.* **456** (1993) 89.
21. W. Kaminsky, C. Piel, K. Scharlach, *Macromol. Symp.* **226** (2005) 25.
22. J. A. Ewen, R. L. Jones, A. Razavi, J. P. Ferrara, *J. Am. Chem. Soc.* **110** (1988) 6255.
23. N. Ishihara, M. Kuromoto, M. Uol, *Macromolecules* **21** (1988) 3356.
24. G. W. Coates, R. M. Waymouth, *J. Am. Chem. Soc.* **113** (1991) 6270.
25. W. Kaminsky, R. Spiehl, *Makromol. Chem.* **190** (1989) 515.
26. W. Kaminsky, A. Ahlers, N. Möller-Lindenhof, *Angew. Chem.* **101** (1989) 1304; *Angew. Chem., Int. Ed. Eng.* **28** (1989) 1216.
27. W. Kaminsky, H. Zielonka, *Polym. Adv. Technol.* **4** (1993) 415.
28. W. Kaminsky, R. Engehausen, K. Zoumis, W. Spaleck, J. Rohrmann, *Macromol. Chem.* **193** (1992) 1643.
29. M. D. Bruce et al., *J. Am. Chem. Soc.* **119** (1997) 11174.
30. C. Piel et al. *Macromol. Chem. Phys.* **207** (2006) 26.
31. W. Spalek et al., *Macromol. Symp.* **89** (1995) 237.
32. A. Torres, K. Swogger, C. Kao, S. Chum, in [3], p. 143.
33. W. Kaminsky, M. Miri, *J. Polym. Sci., Polym. Chem. Ed.* **23** (1985) 2151.
34. T. Seraidaris, W. Kaminsky, J. V. Seppälä, B. Löfgren, *J. Macromol. Chem. Phys.* **206** (2005) 1319.

35. A. K. Hughes, A. J. Kingsley, *J. Chem. Soc., Dalton Trans.* (1997) 4139.
36. C. Cobzaru, S. Deisenhofer, B. Rieger, *Macomol. Chem. Phys.* **206** (2005) 1231.
37. G. Ricci, L. Porri, *Macromol. Chem. Phys.* **198** (1997) 3647.
38. W. Kaminsky, A. Bark, M. Arndt, *Macromol. Symp.* **47** (1991) 83.
39. H. Cherdron, M.-J. Brekner, F. Osan, *Angew. Makromol. Chem.* **223** (1994) 121.
40. N. Herfert, P. Montag, G. Fink, *Makromol. Chem.* **194** (1993) 3167.
41. S. Collins, W. M. Kelly, *Macromolecules* **25** (1992) 233.
42. M. Arndt, I. Beulich, *Macromol. Chem. Phys.* **199** (1988) 1221.
43. A. Zambelli, L. Olivia, C. Pellecchia, *Macromolecules* **22** (1989) 2129.
44. W. Kaminsky et al., *Macromolecules* **30** (1997) 7647.
45. S. Jüngling, R. Mülhaupt, D. Fischer, F. Langhauser, *Angew. Makromol. Chem.* **229** (1995) 93.
46. W. Kaminsky, F. Müller, in M. Terano (ed.): *Current Achievements on Heterogeneous Olefin Polymerization Catalysis*, Sankeisha Co, Nagoya 2004, p. 9
47. K. Soga, M. Kaminaka, *Makromol. Chem. Rapid Commun.* **13** (1992) 221.
48. J. C. W. Chien, D. He, *J. Polym. Sci.: Part A: Polym. Chem.* **29** (1991) 1603.
49. G. G. Hlatky, D. J. Upton, *Macromolecules* **29** (1996) 8019.
50. W. Kaminsky, F. Renner, *Makromol. Chem. Rapid Commun.* **14** (1993) 239.
51. B. L. Booth, G. C. Ofunne, C. Stacey, P. J. Tait, *J. Organomet. Chem.* **315** (1986) 143.
52. T. Kitagawa, T. Uozumi, K. Soga, *Polymer* **3** (1997) 615.
53. J. Saito et. al., *Macromol. Rapid Commun.* **21** (2000) 1333.
54. S. D. Ittel, L. K. Johnson, M. Brookhart, *Chem. Rev.* **100** (2000) 1169.

Microreactors

Wolfgang Ehrfeld, Institut für Mikrotechnik Mainz GmbH, Mainz, Federal Republic of Germany

Volker Hessel, Institut für Mikrotechnik Mainz GmbH, Mainz, Federal Republic of Germany

Verena Haverkamp, Institut für Mikrotechnik Mainz GmbH, Mainz, Federal Republic of Germany

1. Introduction 1
2. Fundamental Advantages of Microreactors 3
3. Microreactor Components 4
3.1. Micromixers 4
3.1.1. Injection of Multiple Microjets 4
3.1.2. Multiple Flow Splitting and Recombination 5
3.1.3. Multilamination of Fluid Layers . . . 6
3.2. Micro Heat Exchangers 8
3.2.1. Cross-Flow Heat Exchange in Stacked Plate Devices 8
3.2.2. Counterflow Heat Exchange in Stacked Plate Devices 10
3.3. Microchannel Catalyst Structures . 12
3.3.1. Partial Oxidation of Propene to Acrolein 13
3.3.2. Selective Partial Hydrogenation of a Cyclic Triene 13
3.3.3. Oxidative Dehydrogenation of Alcohols 14
3.3.4. Gas-Phase Synthesis at High Temperature 15
3.4. Microseparators 15
3.4.1. Exchange between Immiscible Fluids 15
3.4.2. Filtration, Diffusion, and Aerodynamic Separation 17
4. Microreaction Systems 18
4.1. Synthesis of Hazardous Gases in Microfabricated Silicon Devices 18
4.2. Partial Oxidation of Methane in a Stacked Stainless Steel Sheet System 19
4.3. Synthesis of Ethylene Oxide in a System with Integrated Mixing and Catalyst Structures 20
4.4. Oxidation of Ammonia in a Thin Film Silicon Device with Integrated Catalyst, Heaters, and Sensors . . . 22
4.5. Synthesis of a Vitamin Precursor in a Stainless Steel System with Simultaneous Mixing, Reaction and Heat Transfer 23
5. Outlook 25
6. References 26

1. Introduction

In the last two decades powerful processes have been developed for the fabrication of three-dimensional microdevices from a wide variety of materials. These processes include:

- Bulk micromachining of monocrystalline materials, e.g., silicon, by anisotropic wet chemical etching [1]
- Dry etching with low-pressure plasma or ion beams (reactive ion etching, reactive ion beam etching) [2]
- A combination of deep lithography, electroforming, and molding (LIGA process) [3]
- Micromachining with laser radiation [4], [5]
- Anisotropic etching of photosensitive glass [6]
- Advanced mechanical milling, turning, sawing, and drilling processes based on precision engineering [7–11]
- Isotropic wet chemical etching, e.g., of metal foils with a resist pattern [12]
- Micro spark erosion [13], [14]
- Various rapid prototyping processes based on laser-assisted polymerization [15]

Usually, several technologies are combined in a process line for fabricating microdevices and can be regarded as machines in a microworkshop, which also includes auxiliary processes ranging from thin-film technology to mechanical surface modification, as well as test equipment and design and simulation software.

As a result of these microfabrication processes, microminiaturization is no longer exclusively a realm of microelectronics. Mechanical,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b16_b37

optical, thermal, and fluid microdevices are a part of our daily life and include laser heads for CD players, read/write heads for hard disks, microsensors in automobiles, ink jet printer heads, and minimum-invasion surgical instruments [16]. Recently, microdevices have also become of interest for chemical, pharmaceutical, and biotechnical applications [17–28].

In accordance with the term microsystem, which is now widely accepted, microreactors usually are defined as miniaturized reaction systems fabricated, at least partly, by microtechnology and precision engineering. The characteristic dimensions of the internal structures of microreactors such as fluid channels typically range from the sub-micrometer to the submillimeter regime. The terms nanoreactors and milli-/minireactors are sometimes used for devices with characteristic dimensions at the lower and upper limits of this dimensional range. In this article, however, only the term microreactor is used.

Microreactors can be constructed from single components or as integrated systems. Components mainly perform a single operation, mostly a unit operation such as mixing, heat exchange, or separation. The majority of microfluidic devices belong to this group [26]. Since more than one (unit) operation is generally necessary to give the desired effect, many applications require more complex systems, known as integrated microreaction systems. Here components are assembled to form a complete and complex unit, so that macroscopic channels are only found at the end and beginning of the system, while fluid channels in the system have diameters in the micrometer range.

The potential advantages of microreactors compared to macroscopic reaction systems include the following:

- Faster transfer of research results into production
- Earlier start of production at lower costs
- Easier scale-up of production capacity
- Smaller plants for production at distributed sites
- Lower costs for transport, materials, and energy
- More flexible reaction to market demands

Microreaction technology is sometimes regarded as a key strategy for economic growth, by means of cost and time saving, and for the ecology, by sustainable development and saving of natural resources.

In the field of life sciences, commercialization has already been achieved. Several billion compounds have been screened in drug-development programs, and not only 1536-well titer plates but highly sophisticated microsystems for high-throughput synthesis and screening have reached a rapidly expanding market (→ High-Throughput Screening). Immunoassays on integrated optical wave guides and combinatorial DNA analysis and parallel polymerase chain reactions on a chip have become reality.

In chemistry, however, the advantages of miniaturization and the challenge of microreaction engineering have only recently been recognized and inevitably the information is rather limited. The first prototype microreactors for chemical processes were built [20], [27], [28] and tested [20], [29], [30] in the early 1990s. In addition, components of interest for chemical microreaction engineering but developed for other applications were described as early as the 1980s and early 1990s, such as micropumps [31], micromixers [32], micro heat exchangers [33], catalyst supports [17], [33], and miniaturized analytical systems [34], [35], including a gas chromatograph on a silicon wafer [36]. However, systematic and extensive tests of such components in chemical applications were not performed, and straightforward development and fabrication processes for application-specific components and systems for chemical syntheses were lacking.

A new era began in 1995 with the first international workshop on microreaction technology [37], followed in 1997 by an international conference [38], which were initiated and organized by the Institute of Microtechnology Mainz (IMM) and the German Society for Chemical Apparatus, Chemical Engineering, and Biotechnology (DECHEMA) . The development was not only supported by research institutes [19], [21–25] but also by a group at DuPont and BASF [20], [39]. Since then research has begun in the chemical and pharmaceutical industries, and the first microreaction systems are being investigated for commercial applications [39], [40].

2. Fundamental Advantages of Microreactors

It is evident that miniaturization has a number of fundamental advantages in measurement and analysis devices. Miniaturization allows measurements to be performed with a device that needs less space, materials, and energy and often has a shorter response time. Miniaturized devices can be fabricated in parallel by batch processing at low cost and, in particular, a dramatic improvement of performance is achievable by integration of a multitude of miniaturized functional elements, as is well known in microelectronics.

In contrast to measurement systems, which produce information which is independent of the size of the system, reaction systems are designed to produce or convert materials or substances. When reaction systems are miniaturized, the production rate decreases and the value generated by the system per unit time becomes smaller. Therefore, it is usually assumed that the specific production costs increase with decreasing size of the production unit and plant, and it seems to make no sense to replace a large unit by a number of small units.

Nevertheless, in nature substances are always produced in an extremely small production unit, i.e., the living cell, which can be regarded as the most versatile chemical reactor to date. Its small size has been retained over long periods of evolution. To achieve high production rates, nature exploits the advantages of parallel units in the combined action of living cells in multicellular organisms. Besides this fundamental perception supporting the concept of microreaction technology, a variety of technological and commercial arguments have been presented [17], [20], [21], [27], [28], [41–44]. From a general technological viewpoint, the advantages of miniaturized reaction systems can be attributed to two fundamental properties: physical size and number of units.

For given difference in physical properties such as temperature, concentration, density, or pressure, the linear dimensions determine the gradients of these properties. With decreasing linear dimensions, the gradients, and hence the driving forces for heat transfer, mass transport, or diffusional flux per unit volume or unit area, increase. In addition, the surface area to volume ratio increases with decreasing linear dimensions, and therefore the effective exchange surface for mass and heat transfer. The volume of the system and its material holdup decrease with the third power of the linear dimension, and this, in combination with the higher effective exchange surface and gradients, results in shorter response times and should lead to simpler process control.

Because of the intensification of transport properties and the decreased holdup and response time, a variety of process conditions can be applied in microreactors which can not be realized in standard macroscopic devices. A typical example is the realization of nearly isothermal boundary conditions in a highly exothermic reaction by using the extreme heat-transfer rates of microreaction systems. Another example is the speed of mixing, which increases due to shortening of the diffusional path. As a result, the boundary conditions for a chemical reaction can be set over an extremely wide operating range and under favorably uniform conditions usually not achievable with standard laboratory equipment or production plants. Therefore, a major field of application of microreactors is process development to find optimum conditions for a production plant with regard to yield and selectivity [39], [40].

All of these characteristics that result from smaller physical dimensions remain largely unchanged when several miniaturized systems are operated in parallel to achieve the desired production rate. In addition, a larger number of systems results in a higher flexibility in adapting production rate to varying demand since, at least in principle, a certain number of systems may be switched off or further systems may be simply added to the production plant. A plant design based on a large number of small reaction systems can, again at least in principle, be modified to perform a variety of reactions by changing the manner in which the microreactor modules are interconnected. This flexibility may be supported by the considerably broader range of operating conditions of a microreactor compared to a macroscopic system.

Further advantages may result from cost reduction by mass production of modular microreaction systems. To obtain large production rates, many such modules must be assembled. However, there is also the alternative of retaining the

cost advantages of mass production but assembling smaller numbers of modules at various locations. Thus, a distributed production can be realized at tolerable costs, and storage and transport expenditure can be reduced, i.e., the modular approach of microreaction technology should facilitate production on demand and at the point of use.

Because of its small holdup and the flame-arrestor effect of its small channels, a single microreactor can be designed as a inherently safe apparatus. With a proper layout, this inherent safety should be maintainable in a large network of microreactors constituting a production plant. This safety advantage is further increased by the flexibility of distributed production and the short response time for control of a reaction unit. In addition, microreactor-based plants will be designed not with stirred tanks, which are one of the major reasons for operational disturbances, but with micromixing systems that have a negligibly small holdup. As a result, licensing problems might be eliminated, and a shorter time for transfer of laboratory results to production achievable. The possibility of encapsulating a small plant and of avoiding emissions is discussed in connection with small, microreactor-based production plants [41–43].

Major developments in microreaction technology have mainly been made in Europe and the United States [19–21], [23–25], [39]. Industrial and academic research groups are particularly active in Germany and the United Kingdom, but the other European countries are intensifying their development work [37], [38]. Current targets include reactions of explosive, flammable, or highly toxic components and processes with fast interruption of reactions by rapid cooling to suppress side reactions and give products of higher purity. Furthermore, due to advantageous geometries, large exchange rates at interfaces, e.g., in gas-phase catalysis, have been realized, yielding a high degree of conversion. Other developments concern the enhancement of mass transfer by using highly efficient micromixers, the generation of stable and uniform emulsions or dispersions, and the utilization of metastable intermediate species. These broad activities should widen the scope of chemical reactions and process regimes addressable for sustainable development and cost-effective production.

In spite of the scientific, technological, and commercial attractiveness of the microreactor concept in chemistry, it must be emphasized that this discipline is still in its infancy. Up to now, no microreactor process for chemical synthesis has been established commercially and no comprehensive theoretical basis or design rules exist for microchemical engineering. Therefore, this article on microreactors in chemistry should be regarded only as a snapshot of work in progress, which, however, may have a decisive impact in the future.

3. Microreactor Components

3.1. Micromixers

3.1.1. Injection of Multiple Microjets

A basic concept for a micromixer which was used as an integral part of a micro liquid dosing system was described in 1993 [32]. The central element of the mixer is a sievelike structure with a large number of regular holes (Fig. 1). During operation, the mixing area is filled with one liquid, and the other liquid is injected into the mixing volume through a multitude of microholes. Numerous microjets ("microplumes") are generated and increase the contact surface between the two liquids. The holes are positioned in rows 10 – 100 μm apart, which results in short diffusional paths between the jets. Typical flow rates are in the μL/s range, the hole diameter is some 10 μm, and the height of the mixing chamber some 100 μm.

The feasibility of the mixer concept was checked theoretically by numerical analysis of the diffusional flux. It was shown that mixing by injection of many microjets into a fluid stream is much faster than mixing with a large single-point injection (i.e., a flow configuration similar to a mixing tee). Multiple microjets give improved mixing for all flow rates; higher flow rates usually accelerate the mixing process.

Experimental studies included flow visualization for qualitative analysis of the mixing process. Microscopic observation of the jets showed that a homogeneous mixture was established within a few seconds. Absorption measurements were performed directly in the mixing area by using an optical fiber setup. Complete mixing at

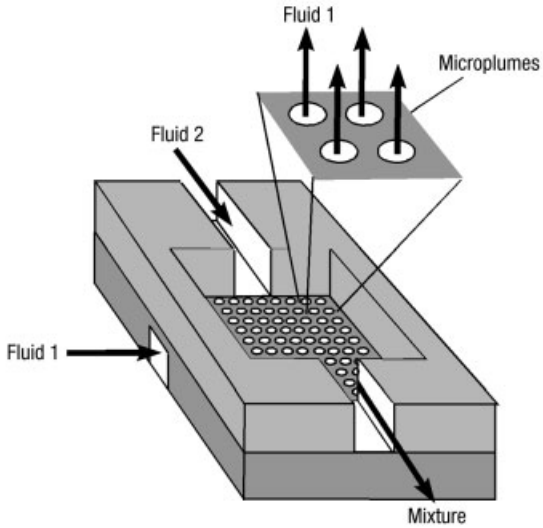

Figure 1. Schematic of a micromixer with injection of multiple microjets into a mixing chamber [32]

a flow rate of about 1 μL/s was found within a mixing time of about 1 s.

3.1.2. Multiple Flow Splitting and Recombination

A micromixer designed by a research group at Danfoss and the Mikroelektronik Centret in Denmark has been used as an integral part of a micromachined chemical analysis system including an industrial chemical sensor [45]. It was operated in the flow range of 0.01 – 0.1 μL/s, which corresponds to a highly viscous flow with a Reynolds number less than 1.

The whole analysis system consists of a silicon/glass sandwich connected by anodic bonding. The mixing structure is shown schematically in Figure 2. One channel structure is etched into glass and the other into silicon. In the region where the channels overlap, they are separated by a structured plate defined by an etch-stop layer. The channel covered by the structured plate was generated by underetching in the [100] direction through slits in the plate. The maximum width of the channels is 300 μm and the depth is 30 μm. The width of the channels for the inlet branches of the micromixer units is about 150 μm. The thickness of the structured plate for separating the channels in the glass and the silicon wafer is 5 μm and the slit width 15 μm.

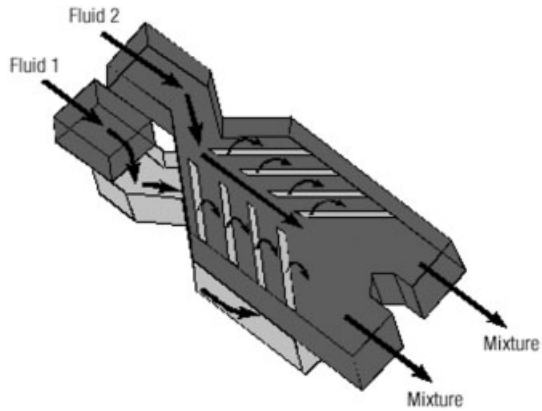

Figure 2. Mixing units of a static micromixer with multiple slit-shaped injection openings [45]

Mixing is performed by penetration of multiple slit-shaped flows into a main stream where the mixing mechanism is similar to that of the concept shown in Figure 1. However, several of the units shown in Figure 2 can be connected in series to improve the mixing quality. This can be of importance if the multiple slit-shaped streams of fluid 1 do not penetrate sufficiently deep into the stream of fluid 2 and contacting layers of the two fluids are formed.

The mixing process was observed by video microscopy, whereby the flow was visualized by of the color change in the reaction of phenol red with an acid. Zones with incomplete mixing

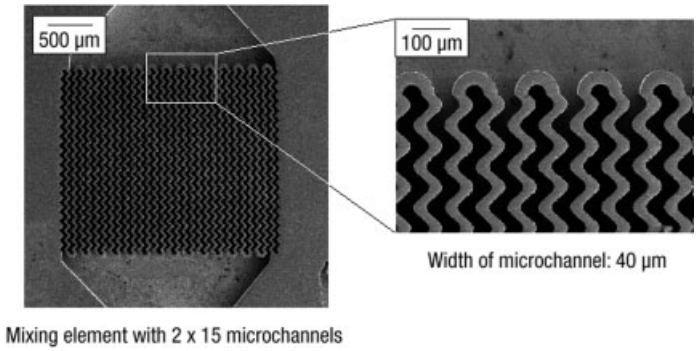

Figure 3. Scanning electron micrographs of a mixing element based on multilamination of thin fluid layers. The device consists of 2 × 15 interdigitated microchannels with corrugated walls, fabricated by LIGA technology [49]

were resolved near the slits. Complete mixing over the cross section of the outlet channel was obtained in 100 – 300 ms.

Micromachining with the Laser LIGA method has been used to fabricate similar devices with multiple flow splitting and recombination [5], [46]. Such mixing configurations have also been realized by wet-chemical anisotropic etching of monocrystalline silicon [47].

3.1.3. Multilamination of Fluid Layers

By means of the LIGA process micromixing devices with various characteristic dimensions have been fabricated which use the multilamination of flow with subsequent diffusional mixing [18], [26], [48], [49]. The design of such a LIGA micromixer is shown in Figure 3. The fluids to be mixed are introduced into the mixing element in counter-flow and stream into an interdigitated channels with corrugated walls. Typical channel widths are 25 or 40 µm. The channel configuration leads to a periodical arrangement of flow lamellae of the two fluids. The lamellar flow leaves the device perpendicular to the direction of the feed flows and, because of the thinness of the lamellae, fast mixing takes place by diffusion (see Fig. 4). The corrugated channel walls increase the contact surface of the lamellar streams and improve the mechanical stability of the separating walls.

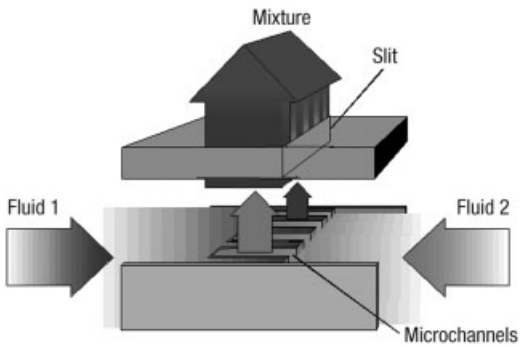

Figure 4. Multilamination of streams in channels with corrugated walls, leading to fast mixing by diffusion [48]

The optimum operating conditions of such a device depend on the height and the length of the channels and the properties of the fluids such as viscosity, flow rate, and pressure. The flow and mixing conditions can be adjusted by withdrawing the fluid through a slit which forces interpenetration across a defined contact zone.

Figure 5 shows the assembly of the micromixer. The LIGA device is inserted into the bottom part and covered with a top part which contains the two inlets for the feed streams and the outlet for the mixture. The discharge slit is also integrated in the top part. A wide variety of materials, including plastics, metals, metal alloys, and ceramics, are suitable for the fabrication of interdigitated structures by the LIGA process and its variants. There is practically no restriction concerning the materials of the top and bottom parts. For commercial devices various metals (e.g., nickel and silver) have been favorably applied. The top and bottom parts are

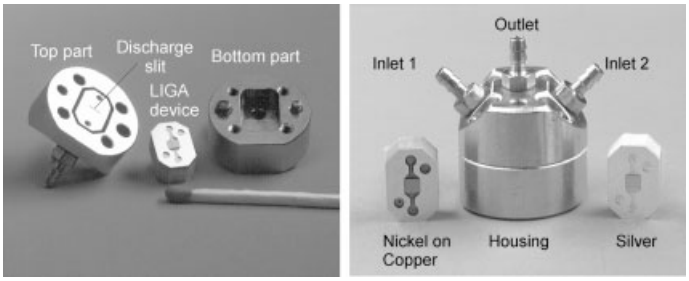

Figure 5. Micromixer consisting of a LIGA device with an interdigitated channel structure and top and bottom parts (Source: IMM [48], [49])

usually made of high-alloy stainless steel and in some cases of plastics.

Micromixing systems of the type shown in Figures 3 and 5 have been tested extensively with miscible fluids. The mixing quality was determined [48] by using feed fluids which react after mixing via two different pathways, a fast one and an ultrafast one, to give differently colored products. In this case the fast reaction gives a colored product, while the result of the ultrafast reaction, which dominates at ultrafast mixing, is a colorless solution. The measurements showed that complete mixing is obtained a few millimeters downstream of the contact zone of the lamellar flows, which corresponds to a mixing time of some 10 ms. The flow rate in these experiments ranged from 10 to 1000 µL/s, and the pressure losses were between 10 and 1000 mbar. For qualitative comparison, the same feed fluids were mixed by using a mixing tee with equal sizes of feed and outlet channels and in a strongly stirred mixing vessel. Such standard devices could not compete with the mixing speed of the LIGA micromixer.

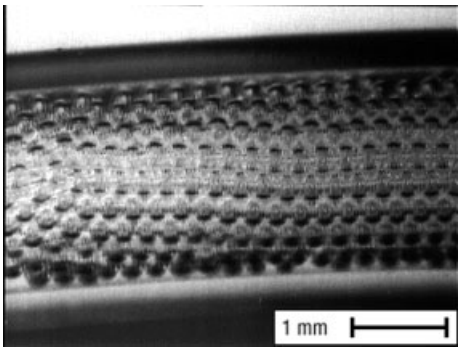

Figure 6. Flowing dispersion of argon bubbles in an aqueous solution generated by means of a LIGA micromixer [50]

Particularly interesting is the utilization of LIGA micromixing systems for generating gas/liquid dispersions [50] and emulsions [51]. Typical results are shown in Figures 6 and 7. Extremely narrow distributions of gas bubbles in liquids and of droplets of one liquid in another can be obtained. These narrow distributions are limited to certain boundary conditions and compositions. The model silicon oil/water emulsion (Fig. 7) was generated without any standard emulgator. Some such emulsions generated by the LIGA micromixer proved to be stable for several months.

Other types of mixing devices using the principle of flow multilamination with subsequent diffusional mixing have been described. Their microstructures differ from the interdigitated channel configurations. Mixers comprising stacked platelets with hundreds of parallel microchannels and a mixing chamber were realized at the Karlsruhe Nuclear Research Centre [52]. Industrial tests were conducted to evaluate their performance in chemical reactions

The two fluids to be mixed are guided in adjacent platelets, which are fed from separate fluid reservoirs. The orientation of the microchannels between reservoirs and mixing chamber is either straight or curved. Thus, an alternating arrangement of fluid layers with a thickness of about 100 µm is generated. Mixing is performed by diffusion between these thin layers within the mixing chamber. The platelets were fabricated by mechanical surface shaping of metal tapes [10]. Various mixers with platelets made of stainless steel, Hastelloy, and palladium were realized.

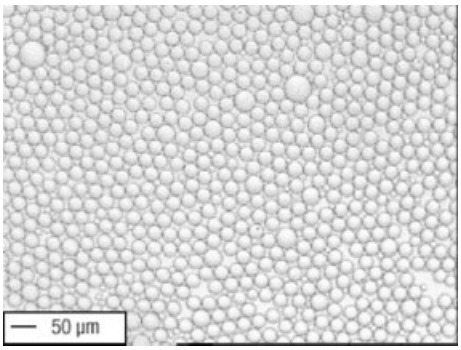

Figure 7. Emulsion of silicon oil droplets in water generated by means of a LIGA micromixer. The emulsion contains no standard emulgator [51].

3.2. Micro Heat Exchangers

To effectively transfer heat from one flowing fluid across a solid boundary to another flowing fluid, sufficiently large contact areas and temperature gradients are required. Another important criterion is the ratio of heat transfer to pressure loss. This is particularly favorable in compact heat exchangers in which the fluid stream is split into many partial streams of small dimensions (e.g., plate heat exchangers). These partial streams are characterized by correspondingly low Reynolds numbers, i.e., they show viscous rather than turbulent behavior. Decreasing the dimensions of the fluid flow increases the temperature gradients and the exchange surface to volume ratio, that is, miniaturization leads to a better heat exchange. The decreased flow dimensions are inevitably associated with increased viscous losses, but the overall heat transfer to pressure loss ratio is improved.

3.2.1. Cross-Flow Heat Exchange in Stacked Plate Devices

This basic concept of performance improvement through miniaturization was demonstrated at the Karlsruhe Nuclear Research Centre, where the development of cross-flow micro heat exchangers was started in the 1980s [10], [11], [53]. A typical cross-flow micro heat exchanger is shown in Figure 8. About 100 platelets, several square centimeters in size and containing rectangular microchannels, are stacked crosswise and bonded hermetically. In this way, two separate passages for the heat transfer fluid and the process fluid with about 4000 microchannels are formed. The cross section of a single microchannel is $100 \times 80\,\mu m$, the material thickness between the two fluids in the crossing channels is $20-25\,\mu m$. The stack is fitted with top and cover plates and connected to fittings for the inlet and outlet ducts of the heat exchanger fluids.

The active volume of such a micro heat exchanger is typically $1\,cm^3$ with an inner surface area of $300\,cm^2$ and a heat transfer surface of $150\,cm^2$. The passages are helium-tight both with respect to each other and to the outside. Because of the small dimensions of the channels and the strong bonding, a relatively high operating pressure can be applied. Such devices have

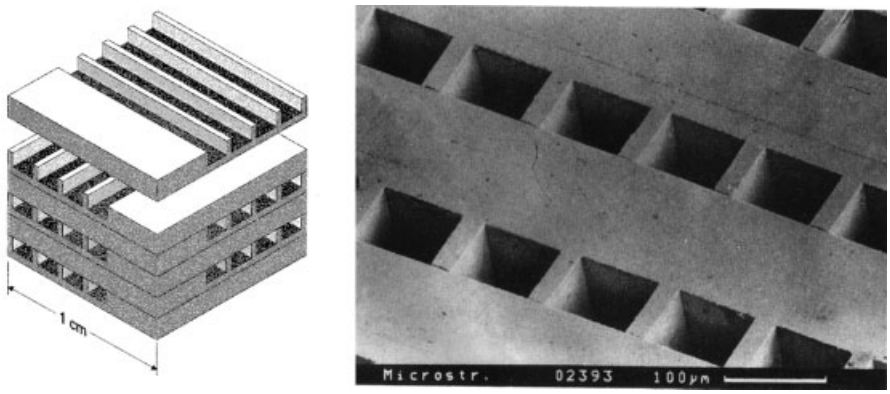

Figure 8. Central component of a cross-flow micro heat exchanger consisting of a stack of crosswise oriented metal platelets (source: Forschungszentrum Karlsruhe)

been tested at pressures up to 25 bar, but this is presumably not the upper limit. Micro heat exchangers have been made from various materials, including aluminum alloys, copper, silver, titanium, and stainless steel. The geometry of the channels, the number of stacked platelets, and the construction of the macroscopic fluid connectors can be adapted to a wide variety of applications [33], [54].

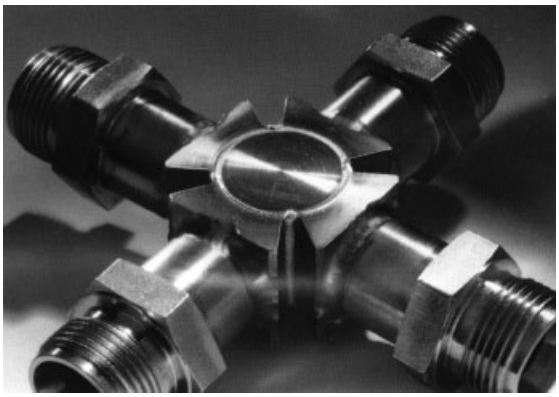

Figure 9. Micro heat exchanger with connections for fluid supply (source: Forschungszentrum Karlsruhe)

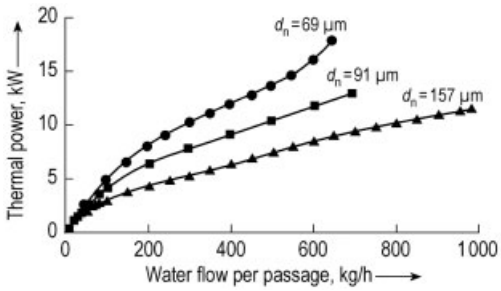

Figure 10. Influence of the hydraulic diameter of the microchannels on the performance of a cross-flow micro heat exchanger, shown by the dependence of heat transfer rate on water mass flow (source: Forschungszentrum Karlsruhe)

Microfabrication of the channels in the platelets is performed by mechanical surface shaping of metal tapes [10]. When aluminum alloys and copper are used, parallel grooves are cut into the tape by ground monocrystalline diamonds. Since iron alloys are incompatible with diamonds, surface shaping of stainless steel is performed with ceramic microtools, but these inevitably result in larger tolerances. With titanium the best results are achieved with polycrystalline diamond tools.

After the grooves have been cut, the tape is cut into platelets of the desired dimensions, which are stacked crosswise to form a cube. The platelets are bonded together by diffusion bonding, and the cube is connected to the macroscopic fittings by electron-beam welding (Fig. 9).

In Figure 10 the measured thermal power (heat transfer rate) is plotted versus the mass flow rate of water per passage for three micro heat exchangers with different hydraulic diameters of the microchannels [52]. The heat transfer rate increases with decreasing hydraulic diameter of the channels. However, since the decreased hydraulic diameter also results in higher pressure losses for a given mass flow and channel length, the ratio of heat transfer to pressure loss is the fundamental performance criterion.

From the results shown in Figure 10 the heat transfer coefficients were calculated (Fig. 11) [52]. Even at the maximum flow rate, laminar flow conditions should exist in each single microchannel. The overall heat transfer coefficients achievable with the micro heat exchanger reach values of more than $20\,\mathrm{kW\,m^{-2}\,K^{-1}}$, which considerably exceed those of macroscopic devices.

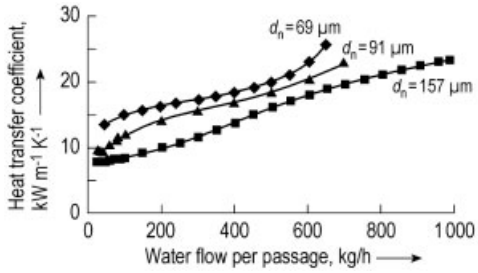

Figure 11. Overall heat transfer coefficients versus mass flow for three heat exchangers with microchannels of various hydraulic diameter (source: Forschungszentrum Karlsruhe)

At a mean temperature difference of 59.3 K, a power of 19.2 kW was transferred in an active volume of $1\,\mathrm{cm^3}$. The overall heat transfer coefficients were up to $25\,\mathrm{kW\,m^{-2}\,K^{-1}}$, corresponding to a volumetric heat transfer coefficient of $0.3\,\mathrm{kW\,cm^{-3}\,K^{-1}}$.

Micro heat exchangers with an active volume of $27\,\mathrm{cm^3}$ and a power of up to 200 kW are now available [52], [55].

3.2.2. Counterflow Heat Exchange in Stacked Plate Devices

The LIGA process can be used to manufacture three-dimensional microstructures of nearly any cross-sectional shape from a wide variety of materials. This flexibility allows counterflow micro heat exchangers to be fabricated with extremely high precision, including inlet and outlet ducts ands other auxiliary structures. Compared to other heat exchange configurations, counterflow is the most efficient from a thermodynamic point of view. The principle of such a micro heat exchanger module is shown in Figure 12 [18], [49]. It consists of platelets with an outer frame and an inner thin membrane, stacked to form a plate-type heat exchanger. The membranes have parallel fins that form channels that guide the fluids and increase the mechanical stability of the device, which is of major importance for operation at high differential pressure. The platelets comprise openings in their corners, where two diagonal openings alternately form the inlet and outlet ducts for one fluid. The other fluid passes through the remaining openings to the adjacent platelet. On both sides of the membranes, which are very thin to ensure efficient heat transfer, uniform countercurrent flow is established by the parallel fins. The total amount of heat transferred in such a stacked device is determined by the number of platelets, which is limited by the pressure losses in the inlet and outlet channels.

A prototype of a micro heat exchanger based on the above principle is shown in Figure 13. It consists of stacked platelets clamped between a top and bottom part made of a high-temperature polyether ether ketone (PEEK) engineering plastic. The top part contains the inlet and outlet ducts for the two fluids. The platelets are 20×20 mm square copper foils with a thickness of 20 to 500 μm. The foils are equipped in their center area with 37 nickel fins with a width of 30 μm and a height of some 100 μm, which form fluid channels with a width of 300 μm. The heat transfer area is to 140 mm^2 per platelet. Each foil is tightly connected to a nickel frame whose thickness corresponds to the height of the fins. The frame also includes the openings of the fluid ducts and alignment structures.

The platelets of the micro heat exchanger shown in Figure 13 were made by the LIGA process, i.e., deep lithography and subsequent electroforming. In the first step, the complementary pattern of the fins and the frame is generated by X-ray lithography in a thick resist layer deposited on an electrically conducting substrate, usually titanium. In the second step, nickel is electrodeposited into the gaps of the lithographic structure to form the fins and the frame. Then, a thin copper film is sputtered onto the resist layer, the fins, and the frame. A thick copper coating is grown on the sputtered copper film by electrodeposition. The final thickness of the copper foil is adjusted by mechanical posttreatment. The duct openings are made by drilling.

Since the above fabrication method is not suitable for mass production, a cheap replication method was developed. It starts with a platelet of a micro heat exchanger manufactured by deep lithography and electroforming as described above. The platelet is then used as an electrode to transfer its complementary structure into a hard metal by micro electro discharge machining (LIGA-μ-EDM). In this way, a precise tool of high mechanical stability is obtained which allows cheap mass production of platelets by embossing. Promising results have been obtained with relatively soft materials like aluminum. This embossing process is being extended to stainless steel and other materials. Powder injection molding (PIM), punching, and wet etching are also under development for cost-effective and application-specific mass fabrication of micro heat exchangers and other components of microreactors.

The performance of the counterflow micro heat exchanger was demonstrated in preliminary measurements of heat transfer between streams of cold (20 °C) and hot (75 °C) water. Measurements of heat transfer rates as function of volume flow rate were performed on four-layer stacks of two types of platelets: one with a 450 μm thick fluid layer, and the other with a fluid layer thickness of 100 μm (Fig. 14).

The heat transfer rate increases with decreasing thickness of the fluid layer. The dependence of the heat transfer coefficient on the volume flow rate, calculated from the data in Figure 14 is shown in Figure 15. The heat transfer coefficients achieved in these preliminary measurements considerably exceed those of conventional heat exchangers.

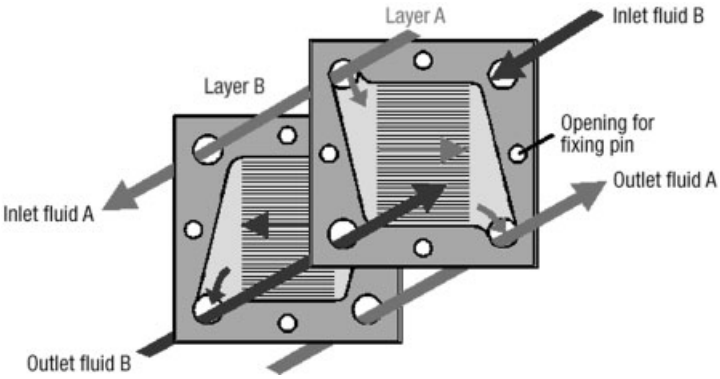

Figure 12. Schematic of the design and flow configuration of a plate-type counterflow micro heat exchanger [49]

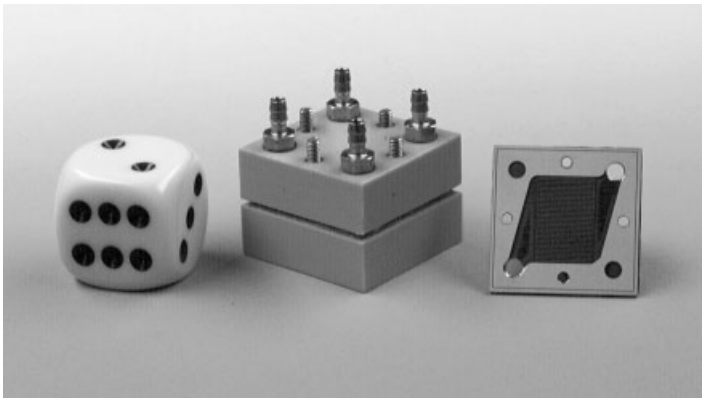

Figure 13. Assembled counter-flow micro heat exchanger with PEEK housing and single platelet [49]

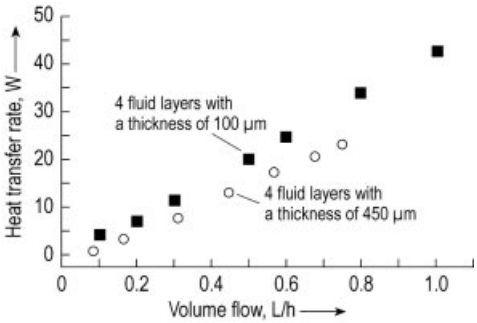

Figure 14. Heat transfer rate versus volume flow rate for a heat exchanger with four platelets of 100 or 450 µm thickness [51]

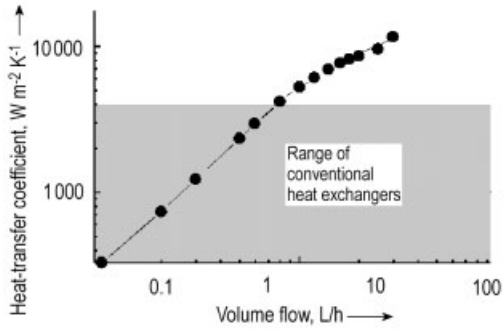

Figure 15. Increase of heat transfer coefficient with volume flow for a counterflow micro heat exchanger comprising four platelets with a fluid layer thickness of 100 µm. The performance range of conventional heat exchangers is also shown [51]

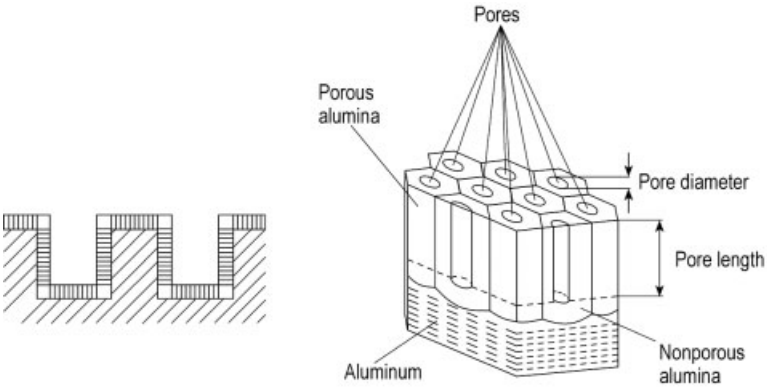

Figure 16. Schematic of the cross section of a microstructured aluminum platelet with anodically formed oxide layer (left) and nanostructure of the pores in the oxide (right) [56]

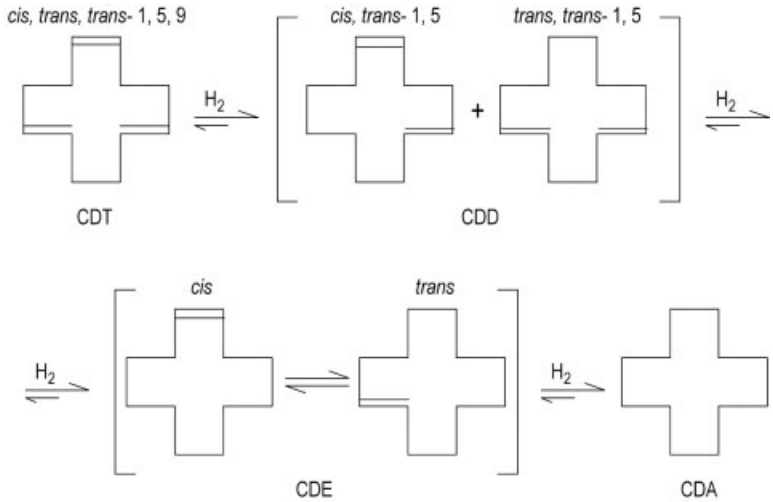

Figure 17. Reaction route of the hydrogenation of *cis*,*trans*,*trans*-1,5,9-cyclododecatriene [29]

3.3. Microchannel Catalyst Structures

Catalysts are incorporated in microreactors as microstructures consisting of bulk catalytic material by wet impregnation of nanoporous supports, physical vapor deposition (PVD), chemical vapor deposition (CVD) of catalytic thin films on microstructured surfaces, and by packing of channels or other microstructures with fine-grained catalytically active powder. The cross-flow micro heat exchanger described in Section 3.2.1 was also used to perform catalytic gas-phase reactions (see Sections 3.2.1 and 3.2.2) [52]. The materials for the channeled platelets included aluminum, copper, silver, and stainless steel. In the case of aluminum platelets, anodic oxidation was used to create an oxide layer with a regular pore structure in which palladium was embedded as a catalytically active material [29], [56–58]. Copper platelets with surface oxidation [24], [30] and silver platelets [28], [40] served as bulk catalysts. For high-temperature applications, platinum ribbons were used with microholes generated by micro electro discharge machining [19]. To investigate controlled catalytic combustion of explosive gas mixtures, stainless steel platelets with channels were covered with Al_2O_3 by CVD, and platinum was applied as catalytically active material by wet impregnation [59].

3.3.1. Partial Oxidation of Propene to Acrolein

One of the first reactions investigated in the flow-through configuration in stacked platelets with microchannels was the heterogeneously catalyzed partial oxidation of propene to acrolein.

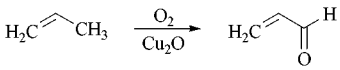

The microchannel reactor used for these experiments, which was developed at the Forschungszentrum Karlsruhe, consisted of copper with a catalytically active surface of Cu_2O [24], [30]. Typical reaction conditions are a temperature of 350 – 375 °C and concentrations of propene and oxygen of less than 1 vol % in nitrogen as carrier gas. The propene conversion rate and the selectivity changed considerably during long-term operation of the reactor, and this is attributed to a change in the composition of the catalytically active layer. Using bromomethane ($c = 2.25$ ppm) as a promoter to stabilize the metal oxide layer gave an increase in selectivity and a slight decrease in conversion.

3.3.2. Selective Partial Hydrogenation of a Cyclic Triene

Compared to macroscopic systems, microchannel devices have a much larger surface area to volume ratio. However, microchannel devices never reach the values of the nanoporous structures typical of many catalysts and catalyst supports. The relatively low conversion rate in the partial oxidation of propene to acrolein (see Section 3.3.1) was attributed to the small catalytic surface area of the microchannels consisting of nonporous cuprous oxide.

Therefore, a method was developed for the generation of a highly porous surface on the platelets and microchannels. The surface of microstructured aluminum platelets was modified by anodic oxidation in aqueous oxalic acid to give a thin layer of Al_2O_3 with a regular structure of pores, the depth of which can be adjusted by means of the operating conditions of the electrochemical process. This porous nanostructure served as a support in which catalytically active components could be embedded [24]. Figure 16 shows a schematic of the channel microstructure and the oxide nanostructure. The distribution of the pore diameters, which are in the range of some tens of nanometers, is highly uniform, a fact which long been utilized for the colorization of anodized aluminum and for the production of isoporous alumina filtration membranes. The surface area to volume ratio is thus increased by a factor of 10^4 to 10^5, and values comparable to those of fixed-bed catalysts are attainable. However, the regular distribution of pores of almost equal size and depth is a clear advantage of the microreactor compared to standard catalyst beds.

The performance of this microreactor was tested in the stepwise hydrogenation of *cis*,*trans*,*trans*-1,5,9-cyclododecatriene to the partially hydrogenated, thermodynamically unstable intermediate cyclododecene (CDE) with the aim of achieving as high a yield as possible (Fig. 17) [29], [55], [56], [57]. For comparison, other catalysts such as conventionally coated granules, pieces of activated aluminum wire, and pieces of activated aluminum foil were used. Palladium was applied as catalyst.

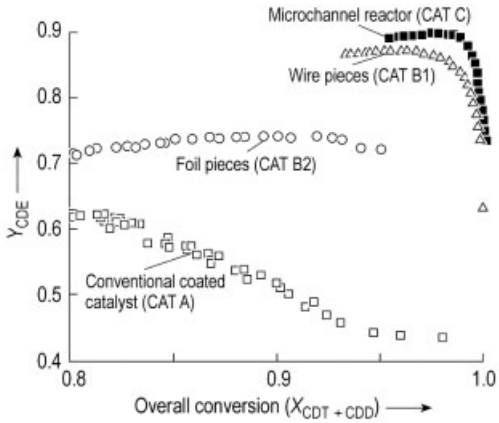

Figure 18. Yield and conversion enhancement by highly ordered catalyst configuration and flow conditions. The microchannel reactor achieves much better values than a catalyst bed with coated granules; partially ordered configurations give intermediate values (CAT A: pore depth 240 µm, commercial industrial palladium catalyst; CAT B1, B2, C: pore depth 37 µm, wet impregnation of anodic oxidized aluminum with palladium acetyl acetonate precursor) [29], [55]

The microchannel reactor proved to have clear advantages. Compared to standard devices

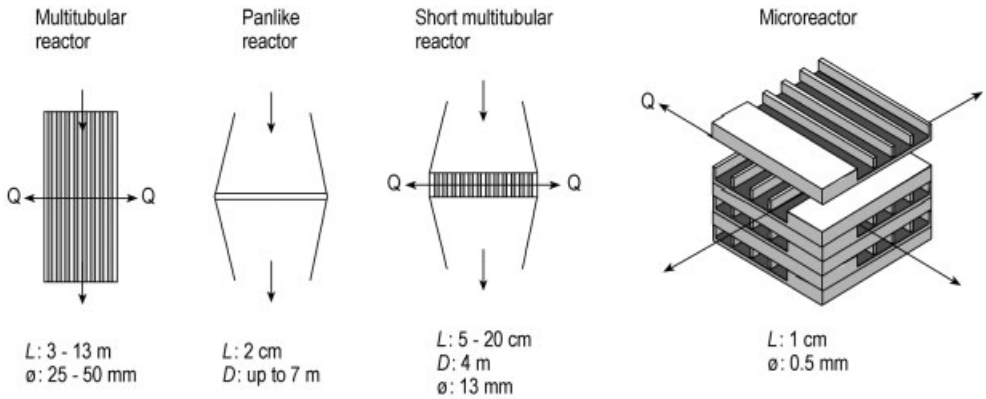

Figure 19. Reactor configurations for gas-phase synthesis of aldehydes by oxidative dehydrogenation of alcohols [44]

with conventionally coated granules it gave a 25 % higher yield of CDE. This was attributed to the regular pore system with defined and relatively short depths of the pores, the uniform distribution of the catalytically active component in the pores, and the uniform flow conditions in the microchannels. Structures which are more regular than coated granules such as pieces of activated aluminum wire or foil gave results closer to those of the microreactor (Fig. 18).

3.3.3. Oxidative Dehydrogenation of Alcohols

Aldehydes are mostly produced by oxidative dehydrogenation of alcohols. About 100 years ago a gas-phase process for the production of formaldehyde on a silver catalyst was developed at BASF to produce formaldehyde (R = H) [40], [44].

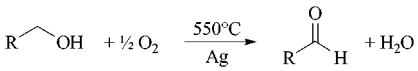

Almost quantitative conversion with a selectivity of 95 % was achieved on a small scale on a large scale in a reactor of diameter 7 m.

Researchers at BASF tried to establish a similar process for the synthesis of derivatized aldehydes. In a small-scale pan-shaped reactor (diameter 5 cm) a selectivity of 90 % at 50 % conversion was achieved, but the selectivity dropped to 45 % when the reactor diameter was increased to 3 m. Therefore, this process was regarded as technically infeasible.

It turned out that optimum conditions are attained in an isothermal process with a short residence time. This explains the unsuitability of multitubular reactors with a tube length of several meters, which allows isothermal operation but has a long residence time. Considerable improvements were obtained with a short multitubular reactor (tube length ca. 10 cm; see Fig. 19).

By means of a microchannel reactor, developed at the Forschungszentrum Karlsruhe [28], a cross-flow heat exchanger design (see Section 3.2.1) comprising silver platelets with 200 microchannels (cross section $320 \times 400\,\mu m$) a further improvement was obtained. Due to the excellent heat-transfer performance and the small dimensions, isothermal operation is combined with short residence time. The residence time is an order of magnitude lower and the heat-transfer coefficient an order of magnitude higher compared to the short multitubular reactor.

In Figure 20 the selectivity of the oxidative dehydrogenation is shown as a function of the conversion. The short multitubular reactor which was optimized over a period of several years reaches a selectivity of 85 % at a conversion of 50 %. At even higher conversions selectivities of up to 96 % were achieved with the microchannel reactor. This is a promising result, also from a commercial point of view, since the reactants are utilized more effectively and a decrease of the amount of sideproducts will lower

the expenditure for subsequent separation and purification.

BASF researchers emphasize that such developments could not have been achieved by using conventional laboratory techniques, but were only feasible by means of microreactors. They expect a significant reduction of development periods and considerable process improvements by using microreactors as a versatile development tool for optimizing production-scale syntheses.

3.3.4. Gas-Phase Synthesis at High Temperature

In a multilateral industrial project dealing with the production of the hazardous gas hydrogen cyanide by oxidative dehydrogenation of methane and ammonia, researchers at IMM developed a multichannel catalyst device [19]. Uniform heat distribution to realize an isothermal process was again one of the major objectives.

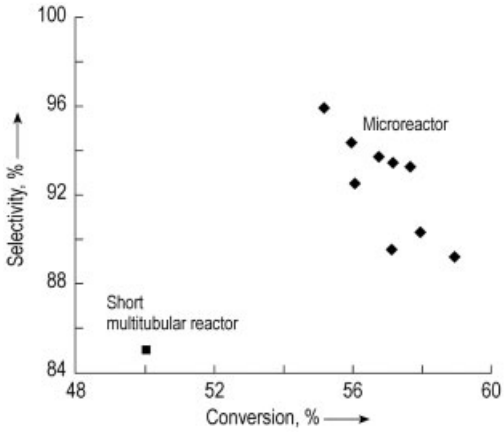

Figure 20. Comparison of the performance in aldehyde synthesis for a microchannel reactor and a short multitubular reactor [40]

The multichannel catalyst device consists of a platinum microstructure embedded in a ceramic support (Fig. 21). The platinum microstructure has parallel microchannels with approximate dimensions of some ten micrometers in diameter and with a length of some hundred micrometers. It can be heated electrically to 1500 K. Structuring of the platinum was performed by micro electro discharge machining, including wire spark erosion and die sinking. The ceramic support was manufactured by precision machining and casting.

Figure 22 shows the catalyst structure glowing in air. The thermal insulation of the hot catalyst structure by the ceramic support is so efficient that the outer border of the ceramic support can be kept close to room temperature.

First experiments prove the feasibility of this catalyst concept for high-temperature reactions. Significant amounts of hydrogen cyanide were measured corresponding to a yield of 32 %, which is about half the value of the industrial process. The conversion of methane is nearly quantitative, however, the selectivity to hydrogen cyanide still has to be improved.

3.4. Microseparators

A number of current investigations in microreaction technology deal with miniaturized separation devices using extraction, filtration, and diffusion processes. The fabrication methods and materials for such devices are described in the preceding chapters and are not discussed in detail here. There are many publications concerning electrophoretic and chromatographic separation devices and miniaturized distillation apparatus. Since these devices are mostly regarded as components of analytical systems, they are not described in this chapter. The reader is referred to [34].

3.4.1. Exchange between Immiscible Fluids

Extraction processes considered here are based on the contact of two immiscible fluids and solute transfer between the two phases. Since miniaturization leads to an increase of surface area to volume ratio, it results in a corresponding enlargement of the exchange interface; hence, intensification of the process by miniaturization should be possible.

A configuration applicable at small dimensions consists of partially overlapping microchannels, which facilitates a largely undisturbed flow of the two phases in their channels and solute exchange between the nondispersed fluids (Fig. 23) [60], [61]. In principle, cocurrent flow, countercurrent flow, cross-flow and other

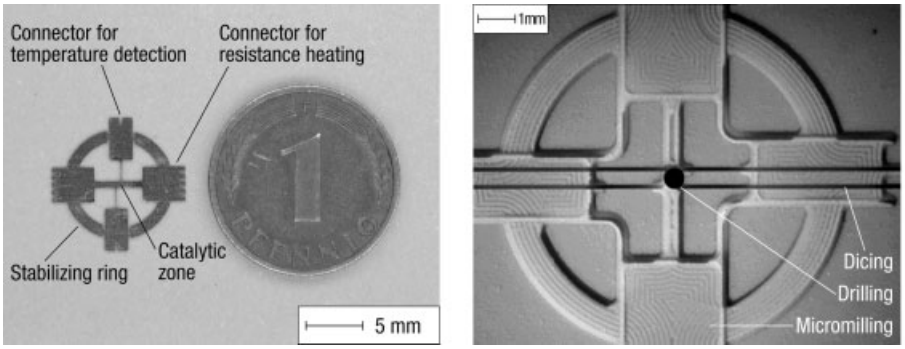

Figure 21. Multichannel device consisting of a platinum catalyst structure (left) and central part of ceramic support made of aluminum oxide (right) [19]

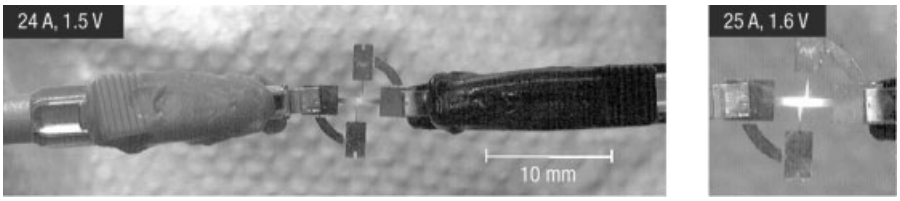

Figure 22. Glowing catalyst structure in air [19]

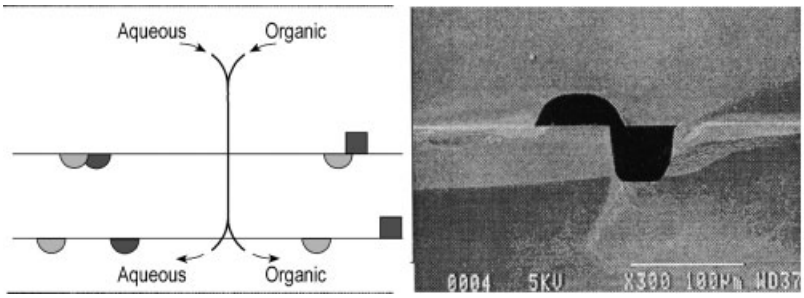

Figure 23. Schematic of solute exchange between immiscible fluids in partially overlapping microchannels (left) and scanning electron micrograph of the cross section of the partially overlapping microchannels (right) [60]

configurations can be realized. Adjustable devices with partially overlapping channels allow the exchange area to be adjusted to the operating conditions.

Further investigations deal with the stability of parallel contacting streams of two immiscible liquid phases which, after solute transfer, are split at the end of the contact zone by wedge-shaped flow divider [62], [63]. Stable flow can be achieved within a certain range of flow rates and viscosities. The stability of the flow is influenced in particular by surface forces, whereas other parameters such as buoyancy, momentum, and viscous stresses are of minor importance. Splitting of the contacting fluids by a wedge-shaped flow divider can be performed with high precision and only minor mixing of the two phases.

Other microextraction units use sieve-like structures to separate the two nondispersed phases [64]. This can be a perforated sheet or channels similar to the device shown in Figure 1. A variant of this type of extraction units has adjacent channels for the two fluids with slits for exchange between the two phases (Fig. 24). In a counterflow device the pressure drop in contacting channels is in the opposite direction and, as a result, a bursting of the fluid from one channel into the adjacent channel may occur. Such a bursting can be avoided by proper adjustment of the slit width.

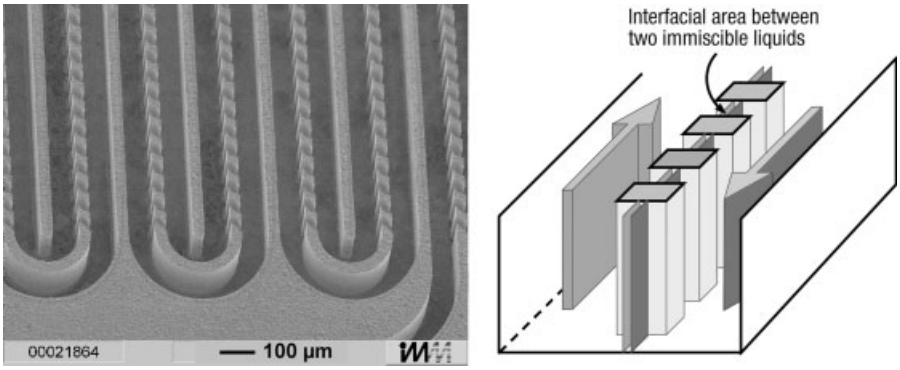

Figure 24. Scanning electron micrograph (left) and schematic (right) of an extraction unit with adjacent channels for two fluids with slits, oblique to the flow direction, for exchange between the two phases [51]

A very effective means for extraction with dispersed phases may be emulsification in micromixers (see Section 3.1.3). Because of the small size of the droplets and the narrow size distribution, a large exchange surface and a sufficient stability may be obtained [51]. Such devices may become of major importance for separation processes based on the mixer/settler principle.

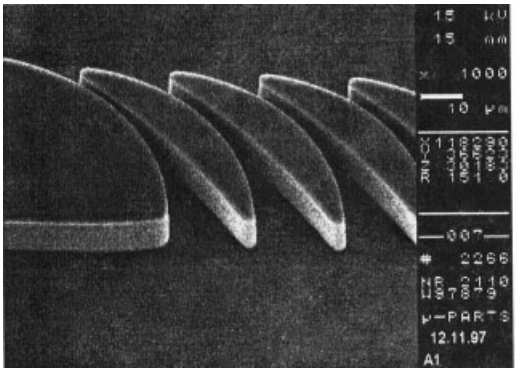

Figure 25. Scanning electron micrograph of a cross-flow filter consisting of lamellae arranged at an angle of attack to the flow direction [65]

3.4.2. Filtration, Diffusion, and Aerodynamic Separation

In the macroscopic range, filtration and sieve structures are often carefully designed with regard to the shape and position of the openings. In the microscopic range usually porous materials with irregular pattern are applied. Microfabrication methods allow the production of completely isoporous microfilters from a wide variety of materials, whereby the size, shape, and position of each pore can be designed. Typical microfilters have pore dimensions in the micrometer range. Special configurations allow the realization of cross-flow filters for concentrating suspended particles or cells. Such cross-flow filters consist of lamellae arranged at an appropriate angle of attack to the flow direction, resulting in a configuration similar to a Venetian blind (Fig. 25) [65].

In the case of membrane units, microfabricated devices are useful as carrier structures with integrated inlets and outlets for fluids. Integrated membrane units are being developed for gas-phase microreactors [66], and microfabricated carrier structures could become important in fuel cells.

At the Karlsruhe Nuclear Research Centre aerodynamic separation systems were developed for the enrichment of the light uranium isotope ^{235}U by means of miniaturized curved nozzles [67]. Analogous to a gas centrifuge, the driving force for the separation of molecules with different masses is the pressure diffusion effect. In spite of a very large development expenditure, the commercial implementation of this so-called separation nozzle method failed, but the LIGA process [3] and the mechanical microfabrication process [27] developed in connection with this method have since become powerful tools for producing a wide variety of microreactor components.

4. Microreaction Systems

4.1. Synthesis of Hazardous Gases in Microfabricated Silicon Devices

DuPont has described the performance of a number of reactions (Table 1) in microfabricated silicon devices [20]. This pioneering development work included detailed analysis of the advantages and applications of microreactors. The devices were not described in detail, but the major components, operations, and the reactions performed have been listed. A modular approach was used to adapt microreaction systems to a variety of reactions, mainly catalytic and noncatalytic gas-phase reactions of hazardous substances. The DuPont microreactor system is based on a stacked silicon wafer architecture consisting of, for example, a static mixer and a reaction chamber for rapid mixing and short contact times. Another type of a microreactor is equipped with a reaction chamber with a "window" for photocatalysis at a specific wavelength.

The formation of the hazardous compound methyl isocyanate has been described in detail. It is synthesized by an exothermic process at high temperatures.

$$[\text{Reactant}]_{\text{Hexane}} + \text{conc. } H_2SO_4 \longrightarrow \text{Byproduct 1}$$

$$\downarrow$$

$$[\text{Intermediate}]_{H_2SO_4} \longrightarrow \text{Byproduct 2}$$

$$\downarrow$$

$$[\text{Product}]_{H_2SO_4} \longrightarrow \text{Byproduct 3}$$

The reaction was run in a microreactor consisting of three etched silicon wafers with a microreactor pattern and two capping wafers (Fig. 26). The reactants were mixed before entering the microfabricated heat exchanger channel packed with polycrystalline silver particles. The product stream was analyzed by capillary gas chromatography.

The experiments successfully demonstrated the feasibility of a microreactor for processing hazardous substances. The conversions for isocyanate formation were in the range of 85 – 95 % for feed conditions and flow rates corresponding to those of a standard laboratory-scale reactor. However, the selectivity did not reach the standard values. The absence of corrosion or material-compatibility problems in the experiments may result from passivation of silicon by oxygen.

Table 1. Selected reactions tested in microfabricated silicon devices [20]

Reactants	Main products	Reaction category
Cyclohexylamine, phosgene	cyclohexyl isocyanate	high temperature, hazardous
Butylamine, phosgene	butyl isocyanate	high temperature, hazardous
Hydrogen chloride, oxygen	chlorine, water	catalytic, hazardous
Carbon monoxide, chlorine	phosgene	catalytic, hazardous
Methylformamide, oxygen	methyl isocyanate	catalytic, high temperature, hazardous
Methane, ammonia	hydrogen cyanide	catalytic, high temperature, hazardous
Dichloromethylsilane, chlorine	trichloromethylsilane, hydrogen chloride	photochemical, hazardous

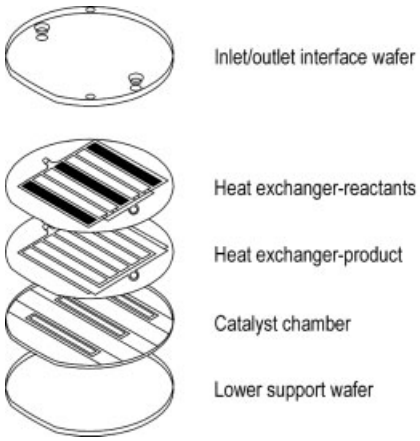

Figure 26. Exploded view of the DuPont microreaction system composed of silicon wafers [20]

Temperature profiles were measured along the catalyst bed and showed a maximum and indicated some heat losses. In spite of the relatively high thermal conductivity of silicon, the high rate of heat generation resulted in a catalyst temperature of up to 300 °C. In contrast to a microreaction system used for a highly exothermic liquid-phase reaction (see Section 4.5 [39], [40]), isothermal operating conditions were not achieved in the DuPont reactor, i.e., the reaction

conditions were between isothermal and adiabatic.

The experiments on the formation of methyl isocyanate in the microreactor clearly demonstrated that microreaction technology is a promising approach for production of hazardous chemicals at the point of use. Although the number of production units increases when a large plant is replaced by distributed microreaction facilities, risks due to transport and storage are eliminated and a smaller total holdup may be realized. Together with the inherent quenching of explosions in microchannels, this should result in a net improvement of safety.

More recent experimental results have demonstrated that more favorable boundary conditions for chemical reactions can be achieved in microreactors compared to macroscopic devices. This may be an increase in selectivity and yield by applying isothermal conditions [39] , [40] or accelerated conversion by using higher temperatures.

4.2. Partial Oxidation of Methane in a Stacked Stainless Steel Sheet System

At the Pacific Northwest National Laboratory (PNNL) of Battelle Institute a microchannel reactor was developed for the oxidation of methane with air yielding synthesis gas (CO, H_2) in a fast exothermic reaction [68].

$$CH_4 + 1/2\,O_2 \rightarrow CO + 2\,H_2$$

This reaction is of interest for generating hydrogen for fuel cells with proton exchange membranes. These fuel cells have a number of advantages compared to high-temperature fuel cells which use ceramic membranes for oxygen ion transport and can be operated with methane.

The hydrogen yield in the partial oxidation of methane is decreased by hot spots in the reaction volume [68]. In addition, thermal runaway effects result in further undesired conversion of the products carbon monoxide and hydrogen. These phenomena can be avoided, in principle, by microreactor systems, which allow isothermal processing and extremely short contact times in the millisecond range. Because of the excellent conditions for heat removal from the exothermic reaction, microreactors should also allow operation at a much higher conversion than macroscopic devices, which usually require dilution with inert gas. In addition, inherent safety may be guaranteed since the small channels and reaction volumes can act like integrated flame arrestors and quench explosions. An excellent catalyst for partial oxidation of methane is rhodium, which inhibits undesired coke formation and hydrogen oxidation because of the high activation energy barrier of rhodium surfaces.

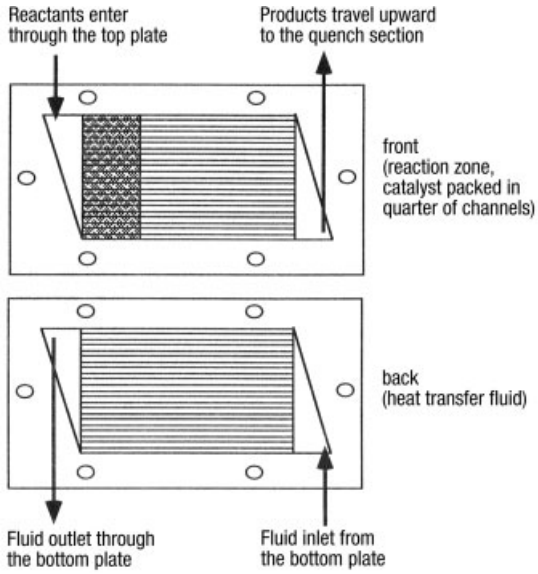

Figure 27. Stainless steel reaction plate [68]

The researchers at PNNL have developed a microreactor for partial oxidation of methane with a planar-sheet architecture. It consists of stacked stainless steel plates, which include combustion, reaction, preheating, quenching, spacer, and top and bottom plates. In the process sequence, the combustion of a fuel is used to heat the reaction volume by means of a gas heat exchanger, and the remaining heat is used for preheating the reactants in a further heat exchanger. The reaction products (H_2, CO) are quenched against air in a third heat exchanger

A schematic of the reaction plate is shown in Figure 27. It has dimensions of $7 \times 3.8 \times 0.4$ cm and contains 37 parallel channels with a depth of 1500 μm and a width of 254 μm separated by fins with a width of 254 μm. The header or the channels over a quarter of their total length are filled with a rhodium-impregnated silica powder. The stainless steel plates are manufactured by preci-

sion engineering or electro discharge machining (EDM). Thin nickel foils are used for sealing the stacked plates.

Extensive tests were performed on the stacked plate microreactor system with a gas chromatograph for analyzing the reaction products. Because of leakage problems the desired operating temperature of 900 °C could not be reached, and the experiments were limited to 700 °C. The inlet feed rate was about 100 $sccm^3$ (standard cubiccentimeter, cm^3 at RT and atmospheric pressure) resulting in a residence time of about 50 ms.

The selectivity and conversion were measured over the temperature range of 300 to 700 °C (channels filled with catalyst) and 300 to 600 °C (header filled with catalyst) and compared to the equilibrium values (Fig. 28). The measured and the equilibrium values differed strongly, as expected for the nonequilibrium process conditions of the experiments. Practically no coke formation was observed, and the concentrations of combustion products (CO_2 and H_2O) were below the equilibrium values. When the front section of the channels was filled with the catalyst, the selectivity for carbon monoxide formation increased with increasing temperature and, in contrast to hydrogen selectivity, exceeded the equilibrium values. When the header was filled with the catalyst, nearly constant selectivities for carbon monoxide and hydrogen over the whole temperature range between 300 and 600 °C were found. This was attributed to much higher local temperatures.

Evidently, the measurements do not cover all aspects of partial oxidation of methane in microreactors, in particular with respect to the local temperature distribution in the reaction zone due to insufficient heat transfer. Nevertheless, these results should be regarded as a first step in the utilization of microreactors in the generation of synthesis gas.

4.3. Synthesis of Ethylene Oxide in a System with Integrated Mixing and Catalyst Structures

On an industrial-scale ethylene oxide is produced with silver catalysts on aluminum oxide having pore sizes of 0.5 – 50 µm at 250 – 300 °C. This reaction is perfectly suited for demonstrating the advantages of microreaction technology. The optimum operating conditions of macroscopic devices and the reaction kinetics are well known, and it is a gas-phase process with a moderate reaction velocity. Ethylene oxide is a volatile toxic substance for which the utilization of microreactors could eventually result in safety advantages. Therefore, a project team including seven research institutes and an advising committee with members from four chemical companies was established to analyze the potential advantages of microreaction systems in this field [19], [69]. The task list of the project team comprises the realization of the microreaction system, including components for mixing, catalysis, gas separation, and analysis, as well as extensive process simulation.

Figure 29 shows the disassembled components for mixing and catalytic reaction of the microreaction system. It was designed and constructed for use under reaction conditions similar to those applied on an industrial scale, i.e., for temperatures up to 300 °C and pressures up to 25 bar. The system comprises stacks of microstructured layers for mixing and catalytic reaction which are inserted in a stainless steel housing. In the process sequence, the reactants ethylene and oxygen are introduced from opposite sides into a static micromixer which feeds a diffusion volume and a subsequent catalyst zone.

The microstructured components for mixing and catalysis were manufactured by a combination of photoablation with excimer laser radiation and electroforming (Laser-LIGA) and by micromilling. The reactor housing was fabricated by conventional milling techniques in combination with micro electro discharge machining and welding [19], [69].

The static micromixer consists of stacked nickel platelets coated with gold to prevent catalytic activity. The platelets have circular concentric channels (Fig. 30, and the inlets to the channels of a platelet are alternately connected to the ethylene and oxygen sides. The thickness of the platelets is 200 µm, and the depth of the channels 50 µm. To obtain equal flow rates and resistances for all channels, the variation in the channel cross sections was determined by a fluidic simulation. The reactants are mixed by multi-entrance feeding into a diffusion volume with a contact length of 1 mm.

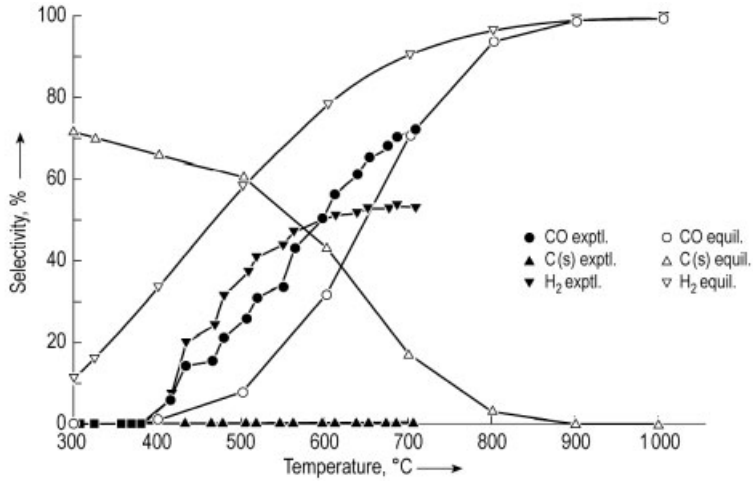

Figure 28. Selectivity for methane partial oxidation: experimental versus equilibrium [68]

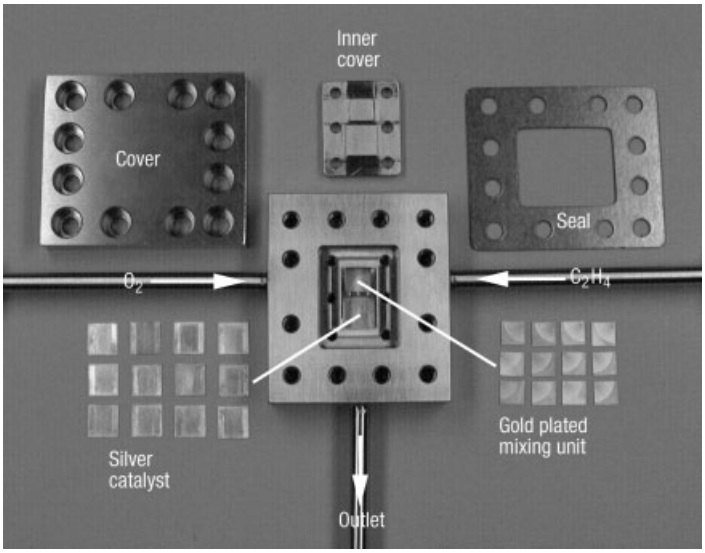

Figure 29. Disassembled microreaction system for ethylene oxide synthesis with housing, fluid connectors, and platelets for stacking catalyst and mixing structures [19]

The catalyst zone following the diffusion volume comprises a stack of solid silver platelets with parallel straight microchannels (Fig. 31).

The channels have a width of 400 µm and a depth of 50 µm to realize short diffusion paths to the catalytic surface. By means of extensive computer simulation it was shown that, even without a heat exchanger, a hot spot of less than 1 K can be achieved and, accordingly, nearly isothermal process conditions for ethylene oxidation exist.

It was found, that the selectivity in the microreactor increases with increasing partial pressure of oxygen. Although characterization of the microreactor is not yet finished, experiments revealed that the performance of the microreactor is close to that of industrial-scale reactors. Without using any chlorine compounds as oxidation inhibitors, a selectivity of 65 % for a conversion of 6 % and a reaction rate of $7 \times 10^{-5}\,\mathrm{mol^{-1}\,m^{-2}}$ has been observed at 350 K and 5 bar, i.e., at low pressure in comparison to the industrial process.

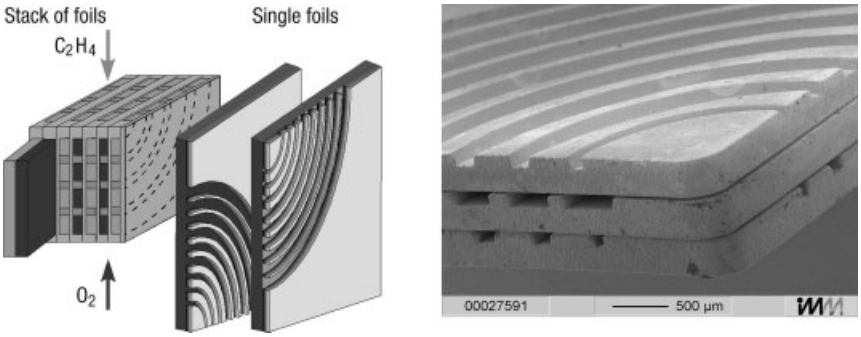

Figure 30. Schematic representation of a static mixer with uniform mass flow consisting of stacked platelets with circular concentric channels (left) and scanning electron micrograph of stacked platelets (right) [69]

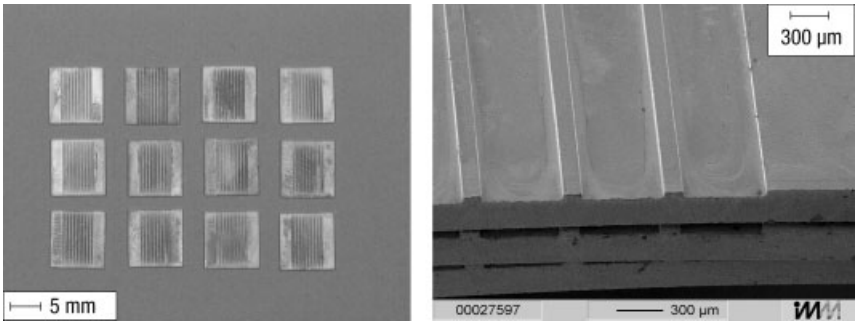

Figure 31. Photograph of silver catalyst platelets with parallel shallow channels (left) and scanning electron micrograph of stacked platelets forming a multichannel catalyst zone [19]

4.4. Oxidation of Ammonia in a Thin Film Silicon Device with Integrated Catalyst, Heaters, and Sensors

The future development of microreactor systems will be based more and more on detailed simulation of the process that takes into account the geometrical boundary conditions of the flow system, as well as thermal, concentration, kinetic, and other data. To provide experimental data and a suitable simulation strategy, a joint research team of MIT and DuPont studied ammonia oxidation in a microreactor. This reaction is particularly suited for such a study because it is highly exothermic and very sensitive to operating conditions such as temperature and composition and cooling or heating effects. It allows an investigation of the ignition and extinction phenomena characteristic of partial oxidation reactions and rapid quenching of reactions.

The simulation was based on a finite-element method and includes a simultaneous solution of momentum, energy, and mass conservation. The geometric boundary conditions were taken fully into account and incorporated three-dimensional simulation, the transport properties in the gaseous mixture, and the kinetic data of the reaction.

Two microreactors were made by means of standard silicon micromachining which mainly differ in their heat removal characteristics [25], [70], [71]. Both types comprise a simple T-shaped channel for introducing the reactants in a counterflow configuration and a longer straight reaction channel (Fig. 32). The T-shaped channel, which has a width and a depth of about 500 µm was etched into a silicon platelet with a size of 15 mm × 25 mm. The bottom of the channel structure was sealed with an aluminum plate with openings for the reactants and the reaction products. The reactants enter at the two inlets of the T-shaped channels and mix at the entrance of the reaction channel. The reaction products leave at the base of the T through the outlet opening in the aluminum plate and are analyzed by a mass spectrometer. On the side oppo-

site to the aluminum plate, the T-shaped channel is covered with a thin membrane which serves as a substrate for deposition of a platinum catalyst layer as well as heaters, temperature sensors, and thermal flow sensors by standard methods of thin-film technology.

In the first reactor type the membrane is made of silicon nitride and has a thickness of 1 µm. Because of the low thermal conductivity of silicon nitride and the small thickness of the membrane, heat generated, e.g., by ignition on the catalyst surface or by the heaters, is removed poorly. The second reactor type has a 2.6 µm silicon membrane which, because of its higher thickness and the much higher thermal conductivity of silicon, removes heat generated near the membrane much more efficiently.

In extensive investigations temperatures were determined as a function of channel width, channel length, time, heater power input, etc. Good agreement between measurements and simulated temperature profiles was obtained for pure oxygen as a nonreacting gas and for the reacting oxygen/ammonia mixture.

An interesting result was due to the different heat dissipation behavior of the two reactors and membranes (Fig. 33). In the silicon nitride case the temperature versus power curve is characterized by a hysteresis loop. With increasing power input the temperature first rises linearly and, at a certain input power, jumps to a much higher value due to ignition. When the power input is reduced the temperature decreases again but initially remains at a higher level and then, at a lower input power, jumps back to the previous curve. Due to its higher heat removal the reactor with the silicon membrane does not show such a hysteresis loop but a continuous increase of temperature with increasing power input.

For given boundary conditions, the different behavior of the two reactors results in a change in product composition of the Pt-catalyzed ammonia oxidation. NH_3 and O_2 are mainly converted to N_2 and NO in the silicon nitride membrane reactor, while significant amounts of N_2O were additionally observed in the silicon membrane reactor.

The investigations clearly show that microreactors allow reaction conditions to be achieved which are not feasible in macroscopic devices. For example, the silicon reactor can be operated in a regime between the ignition and extinction temperatures and give products different from those of a macroscopic device. This change in reaction products may also be of importance in other reactions such as ethylene oxidation and for reactions in the explosive regime in general.

The good agreement between measurement and simulation demonstrates that, at least for relatively simple devices and reactions, a theoretical prediction of the behavior of microreactors and, therefore, detailed optimization is possible.

4.5. Synthesis of a Vitamin Precursor in a Stainless Steel System with Simultaneous Mixing, Reaction and Heat Transfer

For the synthesis of a vitamin precursor a production process was established at BASF using a highly exothermic two-phase liquid/liquid reaction [39], [40], [49]. The two phases consist of hexane and concentrated sulfuric acid. The reactant, initially dissolved in hexane, reacts at the phase boundary to give an intermediate, which is then transferred into the sulfuric acid phase where the product is formed in a further reaction. The reaction is quenched after product formation by addition of water. A major problem of this process is the lowering of the yield by formation of byproducts in thermally induced side and consecutive reactions.

Initially, the synthesis was performed as a semi-batch process with a yield of 70 %. It was later replaced by a continuous process with a mixer and heat exchanger. The residence time in the mixer is 0.2 s and that in the heat exchanger 4 s. In spite of the short residence time, a temperature rise of 35 °C occurs in the mixer. Since this increase in temperature is half of the adiabatic temperature rise of the reaction, 50 % conversion is performed in the mixer. Completion of the reaction proceeds isothermally in the heat exchanger at 50 °C. The yield of the continuous process is 80 – 85 %.

In kinetic investigations with conventional laboratory techniques, the yield decreased dramatically to 25 %. This was attributed to the

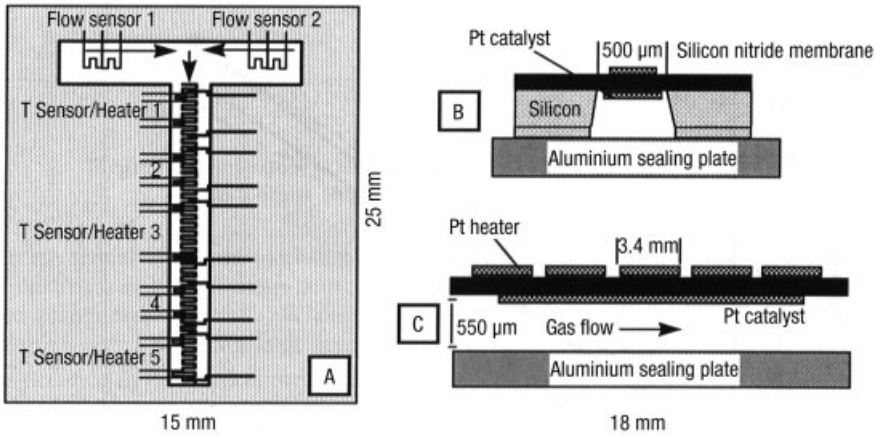

Figure 32. Schematic representation of a T-shaped microreactor for ammonia oxidation manufactured by means of silicon micromachining and thin-film technology
A) Top view; B) Cross section of reaction channel, perpendicular to flow direction; C) Cross section of reaction channel, parallel to flow direction [70]

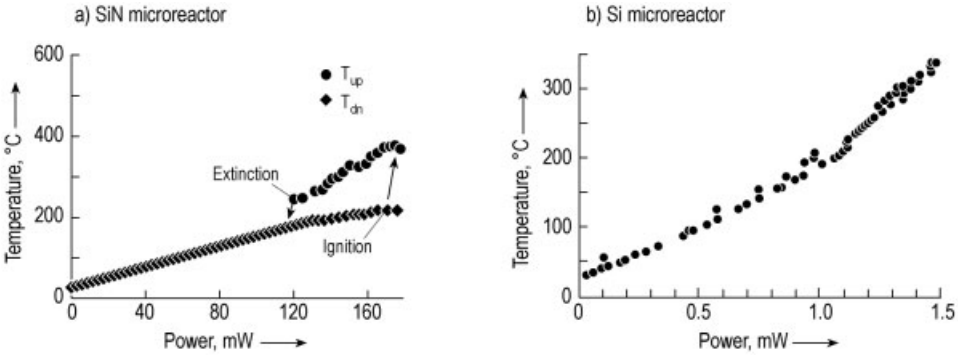

Figure 33. Reaction characteristics for ammonia oxidation in microreactors with low (left) and high (right) heat dissipation near the catalyst zone, demonstrated by power input/temperature curves [72]

fact that the whole procedure, including mixing of two phases, transfer of the heat of reaction, dilution with water, and phase separation, took two minutes. To obtain the desired information about the kinetics, a microreactor was designed and manufactured by IMM that allows the exothermic reaction to be carried out under nearly isothermal conditions [39], [40]. The design specifications were to limit the maximum temperature rise to 1 K and to achieve a total residence time of 4 s.

The five-plate microreactor is shown in Figure 34. Cavities between the bottom plate and the next plate contain four microstructured stainless steel platelets with an integrated mixer and heat exchanger system. The third plate has a delay loop which can be used for varying the residence time between 1 and 10 s. In the fourth plate the reaction is quenched by dilution of the concentrated sulfuric acid with water. This step is even more exothermic than the main reaction, but the product is no longer thermally sensitive under these conditions.

The stainless steel platelets integrating mixer and heat exchanger are the central elements of the reactor and were fabricated by means of a new micro discharge machining process developed at IMM, termed microerosive grinding (see Fig. 35) [73]. Eight heat exchanger channels, 60 µm wide and 900 µm deep, are surrounded by cooling channels machined into the back side of the platelets. Since the reactor con-

tains four of these platelets, the reaction is performed in 32 channels. Mixing is realized by intimate contacting of 32 partial streams of each reactant phase. Thus, mixing and reaction occur simultaneously. Since all volume elements flow along the same distance and hence have the same residence time, no backmixing takes place. The exclusion of backmixing in the microreactor should increase the product yield, since the labile product is sensitive to consecutive reactions.

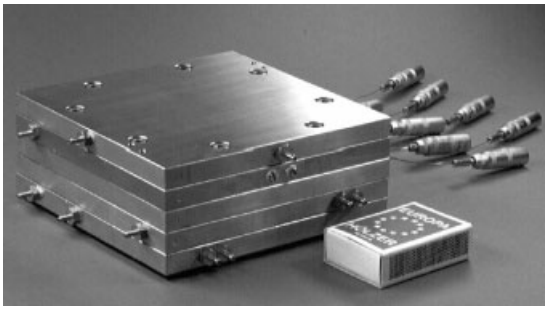

Figure 34. Microreaction system for synthesis of a vitamin precursor comprising integrated static mixers, heat exchangers, reaction channels and delay loop [49], [73]

In the first experiments, the dependency of the yield on the residence time in a microreactor was analyzed at 50 °C (Fig. 36) [39], [40]. For a residence time of 4 s at 50 °C a yield of 80–85 % was found, which corresponds to that of the industrial process. For longer residence times a strong decrease in yield was found. For a residence time of 15 s, the yield decreased to ca. 65 %. This corresponds to the results obtained with conventional laboratory techniques, which inevitably result in long residence times (25 % yield).

In a more detailed optimization of yield and selectivity, the residence time was varied between 1 and 120 s, the operating temperature was decreased to 20 °C, and the exothermic reaction was run isothermally (Fig. 35). At an operating temperature of 20 °C and a residence time of 30 s a maximum yield of nearly 95 % was achieved, which considerably exceeds that of the continuous industrial process. The amount of byproducts is reduced by more than 50 % which clearly demonstrates the capability of microreaction technology.

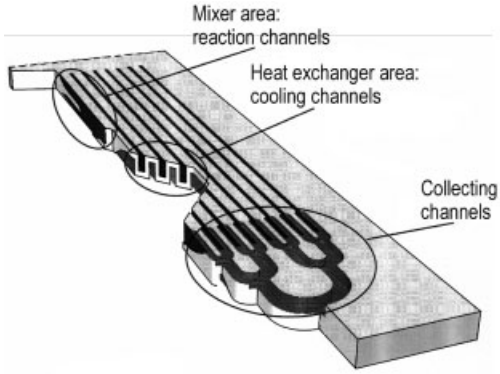

Figure 35. Schematic of a platelet with integrated mixer and heat exchanger [73]

5. Outlook

Although current development work on microreaction technology is promising, it has to be emphasized that, in spite of a rapidly growing number of publications, this technology is still in its infancy. Some potential fields of applications such as electrochemistry [74], photochemistry [17], [20], and electrically steerable electrodes as dynamic catalysts [75] are just at the very beginning. Nevertheless, the most recent results make this technology even more promising, and the first announcements that chemical companies will utilize microreaction components for commercial processing have been made. Further impetus can be expected from developments in related fields.

Parallel organic synthesis of large numbers of compounds, often via complex reaction pathways or in very high purity, will inevitably require microreactor systems. Highly parallel polymerase chain reactions (PCR) are already performed in miniaturized reactors, and, with increasing complexity and automation, the potential growth is large. Progress in the field of microreaction technology in molecular biotechnology is described in [76–85].

A further driving force for the development of microreactor technology are mobile energy systems, especially fuel cells, with the aim of increasing the stored energy density by an order of magnitude. The development work deals not only with automotive applications but also with miniaturized fuel cell power supplies for mobile

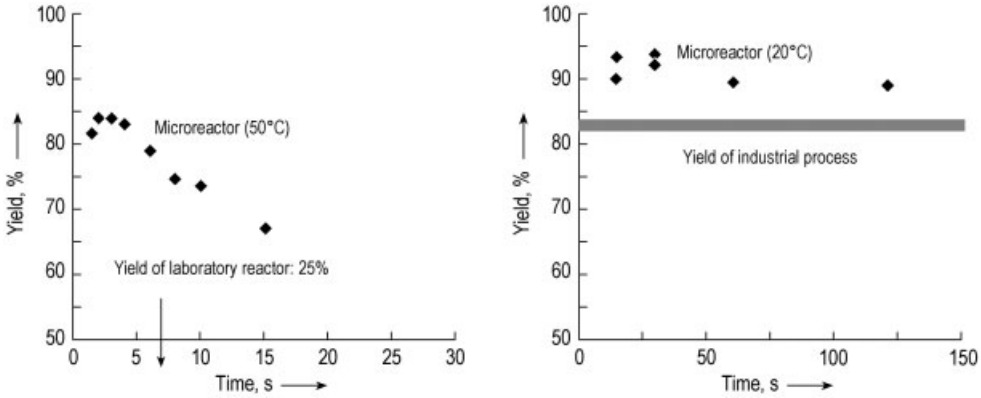

Figure 36. Influence of residence time on the yield of a microreactor for synthesis of a vitamin precursor operated at 50 °C (left) and 20 °C (right) [40]

phones, laptop computers, and other portable electronic equipment. This work is also funded by defense organizations like DARPA in the United States [86], which expect advancement in mobility, safety, and comfort through microreaction technology.

6. References

Specific References

1. K. E. Peterson, *Proc. IEEE* **70** (1982) 420 – 457.
2. R. S. Muller, *Sens. Actuators* **A21** (1990) 1 – 8.
3. W. Ehrfeld, H. Lehr, *Radiat. Phys. Chem.* **45** (1995) 3, 349 – 365.
4. V. Pfeufer, "Präzise Lichtblitze – Mikrobearbeitung mit dem Excimerlaser," *F & M* **104** (1996) 7-8, 532 – 536.
5. J. Arnold, U. Dasbach, W. Ehrfeld, K. Hesch, H. Löwe, "Combination of Excimer Laser Micromachining and Replication Processes Suited for Large Scale Production (Laser-LIGA)," *Appl. Surf. Sci.* **86** (1995) 251 – 258.
6. T. R. Dietrich, W. Ehrfeld, M. Lacher, M. Krämer, B. Speit, "Fabrication Technologies for Microsystems Utilizing Photoetchable Glass," *Microelectron. Eng.* **30** (1996) 497 – 504.
7. S. Steckemetz, R. Hezel, Emmerthal, "Mikrostrukturierung mit der Drahtläppmaschine," *F & M* **106** (1998) 3, 146 – 149.
8. E. M. Kussul, D. A. Rachkovskij, T. N. Baidyk, S. A. Talayev, "Micromechanical Engineering: a Basis for the Low-Cost Manufacturing of Mechanical Microdevices Using Microequipment," *J. Micromech. Microeng.* (1996) 6, 410 – 425.
9. E. Westkämper, H.-W. Hoffmeister, J. Gäbler, "Spanende Mikrofertigung," *F & M* **104** (1996) 7-8, 525 – 530.
10. K. Schubert, W. Bier, G. Linder, D. Seidel, "Profiled Microdiamonds for Producing Microstructures," *Ind. Diamond Rev.* **50** (1990) no. 5.
11. W. Bier, W. Keller, G. Linder, D. Seidel, K. Schubert: *Manufacturing and Testing of Compact Micro Heat Exchangers with High Volumetric Heat Transfer Coefficients,* **DSC-Vol. 19,** ASME, New York 1990.
12. A. Visser, M. Junker, D. Weißinger: *Sprühatzen metallischer WErkstoffe,* Eugen E. Leuze-Verlag, Saulgau 1995.
13. D. F. Dauw, B. Van Coppenolle, "On the Evolutions of EDM research–From Fundamental to Applied Research," in J.-P. S. van Griethuysen, D. Kiritsus (eds.): *Proc. of International Symposium for Electromachining 1995 (ISEM XI),* Presses Polytechniques et Universitaires Romandos, Lausanne, 1995, pp. 133 – 142.
14. W. Ehrfeld et al., "Micro Electro Discharge Machining as Technology in Micromachining," *Proc. of SPIÉs 1996 Symp. on Micromachining and Microfabrication, Conference on Micromachining and*

Microfabrication Process Technology II, **vol. 2879** (1996) pp. 332 – 337.
15. S. Zissi, A. Bertsch, J. Y. Jézéquel, S. Corbel, J. C. André, *Microsystem Technologies* **2** (1996) 97.
16. W. Ehrfeld, V. Hessel, H. Löwe, C. Schulz, L. Weber: "Materials of LIGA Technology," in *Microsystem Technologies,* Fraunhofer IZM, Berlin 1997.
17. W. Ehrfeld, V. Hessel, H. Möbius, Th. Richter, K. Russow: "Potentials and Realization of Microreactors," in W. Ehrfeld (ed.): *Microsystem Technology for Chemical and Biological Microreactors,* **DECHEMA monograph Vol. 132,** VCH Verlagsgesellschaft, Weinheim 1996 pp. 1 – 28.
18. W. Ehrfeld, H. Löwe, V. Hessel, Th. Richter: "Anwendungspotentiale für chemische und biologische Mikroreaktoren," *Chem. Ing. Tech.* **69** (1997) 7, 931 – 934.
19. H. Löwe et al., "Microreactor Concepts for Heterogenous Gas Phase Reactions," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans, LA 1998, pp. 63 – 74.
20. J. J. Lerou et al.: "Microfabricated Minichemical Systems: Technical Feasibility," in W. Ehrfeld (ed.): *Microsystem Technology for Chemical and Biological Microreactors,* **DECHEMA monograph Vol. 132,** VCH Verlagsgesellschaft, Weinheim 1996, pp. 51 – 69.
21. R. W. Wegeng, C. J. Call, M. K. Drost: "Chemical System Miniaturization," *Proceedings of 1996 Spring National Meeting AIChE, Feb. 25 – 29 1996, New Orleans* 1996, pp. 1 – 3.
22. K. Schubert, W. Bier, J. Brandner, M. Fichtner, C. Franz, G. Linder: "Realization and Testing of Microstructure Reactors, Micro Heat Exchangers and Micromixers for Industrial Applications in Chemical Engineering," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans, LA 1998, pp. 88 – 95.
23. J. Shaw, B. Miller, C. Turner, M. Harper, S. Graham: "Mass Transfer of Species in Micro-contactors: CFD Modelling and Experimental Validation," in H. M. Widmer, E. Verpoorte, S. Barnard (eds.): *Proceedings of the 2nd International Symposium on Minaturized Total Analysis Systems, μTAS 96, Basel,* 19 – 22 Nov. 1996, pp. 185 – 188.
24. D. Hönicke, G. Wießmeier: "Heterogeneously Catalyzed Reactions in a Microreactor," in W. Ehrfeld (ed.): *Microsystem Technology for Chemical and Biological Microreactors,* **DECHEMA monograph Vol. 132,** VCH Verlagsgesellschaft, Weinheim 1996, pp. 93 – 107.
25. R. Srinivasan et al., "Micromachined Reactors for Catalytic Partial Oxidation Reactions," *AIChE Journal* **43** (1997) 11, 3059 – 3069.
26. W. Ehrfeld, V. Hessel, H. Lehr: "Microreactors for Chemical Synthesis and Biotechnology–Current Developments and Future Applications," in A. Manz, H. Becker (eds.): *Topics in Current Chemistry, Volume 194: Microsystem Technology in Chemistry and Life Science,* Springer Heidelberg 1998, pp. 233 – 252.
27. W. Bier, G. Linder, K. Schubert: "Mechanische Mikrotechnik," *KfK-Nachrichten,* **vol. 23,** Kernforschungszentrum Karlsruhe, Karlsruhe 1991, pp. 2 – 3, 165 – 173.
28. W. Bier, A. Guber, G. Linder, Th. Schaller, K. Schubert: *Mechanische Mikrotechnik – Verfahren und Anwendungen,* 1. Statuskolloquium des Projekts Mikrosystemtechnik, KfK-Bericht 5238, Kernforschungszentrum Karlsruhe, Karlsruhe 1993, pp. 132 – 137.
29. G. Wießmeier, D. Hönicke: "Strategy for the Development of Micro Channel Reactors for Heterogenously Catalyzed Reactions," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 24 – 32.
30. G. Wießmeier: Diplomarbeit, Universität Karlsruhe, 1992.
31. H. T. G. van Lintel, F. C. M. van den Pol, S. Bouwstra: "A Piezoelectric Micropump Based on Micromachining of Silicon," *Sens. Actuators* **15** (1988) 2, 153 – 167.
32. R. Miyake, T. S. J. Lammerink, M. Elwenspoek, J. H. J. Fluitman, "Micro Mixer with Fast Diffusion," *MEMS-93, Februar 93, Fort Lauderdale, FL,* (1993) 248 – 253.
33. G. Linder, W. Bier, T. Schaller, K. Schubert, D. Seidel: "Mikrowärmeübertrager und Mikroreaktoren," in Tagungsband des Symposiums ″Mikrotechnik″, ACHEMA 94, Frankfurt/M. 1994.

34. A. van den Berg, P. Bergveld (eds.): μ-*TAS: Miniaturized Total Chemical Analysis Systems,* Kluwer Academic Press, Amsterdam 1995.
35. E. M. J. Verpoorte et al., "Three-Dimensional Micro Flow Manifolds for Miniaturized Chemical Analysis Systems," *J. Micromech. Microeng.* **4** (1994) 246.
36. J. B. Angell, S. C. Terry, P. W. Barth, "Mikromechanik aus Silicium," *Spektrum* 1983, June, 38 – 50.
37. Workshop on Microsystem Technology for Chemical and Biological Microreactors, organized by DECHEMA and IMM, Mainz, 1995, see: Microsystem Technology for Chemical and Biological Microreactors, DECHEMA Monographs Vol. 132, W. Ehrfeld, ed., VCH Verlagsgesellschaft, Weinheim 1996, pp. 51 – 69.
38. J. P. Baselt, A. Förster, J. Herrmann, D. Tiebes: "Microreactor Technology: Focusing the German Activites in this Novel and Promising Field of Chemical Process Engineering," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans, LA 1998, pp. 13 – 17.
39. K. P. Jäckel, O. Wörz, "Winzlinge mit großer Zukunft–Mikroreaktoren für die Chemie," *Chem. Tech.* **26** (1997) 1, 130 – 134.
40. O. Wörz, K. P. Jäckel, T. Richter, A. Wolf: "Microreactors, a New Efficient Tool for Optimum Reactor Design," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology,Topical Conference Preprints, March 9 – 12 1998,* New Orleans, LA 1998, pp. 88 – 95.
41. R. S. Benson, J. W. Ponton, "Process Miniaturization–a Route to Total Environmental Acceptability?" *Trans. IChemE.* **71** (1993) A2, 160 – 168.
42. J. W. Ponton: "Observations on Hypothetical Miniaturised, Disposable Chemical Plant," in W. Ehrfeld (ed.): *Microreaction Technology, Proc. of the 1st Int. Conf. on Microreaction Technology,* Springer-Verlag, Berlin 1997, pp. 10 – 19.
43. I. Rinard: "Miniplant Design Methodology," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans, LA 1998, pp. 299 – 312.
44. K. P. Jäckel: "Microtechnology: Application Opportunities in the Chemical Industry," in W. Ehrfeld (ed.): *Microsystem Technology for Chemical and Biological Microreactors,* **DECHEMA monograph Vol. 132,** VCH Verlagsgesellschaft, Weinheim 1996, pp. 20 – 50.
45. J. Branebjerg, P. Gravesen, J. P. Krog, C. R. Nielsen: "Fast Mixing by Lamination," *Proceedings of IEEE MEMS'96, San Diego,* 1996, 441 – 446.
46. V. Hessel, W. Ehrfeld, H. Möbius, T. Richter, K. Russow: "Potentials and Realization of Micro Reactors," in *Proc. Int. Symp. on Microsystems, Intelligent Materials and Robots, Sept. 27th – 29th,* Sendai 1995, pp. 45 – 48.
47. N. Schwesinger, T. Frank, H. Wurmus: "A Modular Microfluid System with an Integrated Micromixer," *J. Micromech. Microeng.* **6** (1996) 99 – 102.
48. W. Ehrfeld, K. Golbig, V. Hessel, H. Löwe, T. Richter: "Characterization of Mixing in Micromixers by a Test Reaction: Single Mixing Units and Mixer Arrays," *Ind. Chem. Res.* (in press).
49. W. Ehrfeld et al.: "Fabrication of Components and Systems for Chemical and Biological Microreactors," in W. Ehrfeld (ed.): *Microreaction Technology, Proc. of the 1st Int. Conf. on Microreaction Technology,* Springer-Verlag, Berlin 1997, pp. 72 – 90.
50. V. Hessel et al.: "Gas/Liquid Dispersion Processes in Micromixers: the Hexagon Flow," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 259 – 266.
51. IMM, unpublished results.
52. K45. Schubert et al.: "Realization and Testing of Microstructure Reactors, Micro Heat Exchangers and Micromixers for Industrial Applications in Chemical Engineering," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 88 – 95.
53. W. Bier et al.: "Gas-to-Gas Heat Transfer in Micro Heat Exchangers," *Chem. Eng. Process.* **32** (1993) no. 1, 33 – 43.
54. Th. Schaller, W. Bier, G. Linder, K. Schubert: *"Mechanische Mikrotechnik für Abformwerkzeuge und Kleinserien,"* 2.

Statuskolloquium des Projekts Mikrosystemtechnik, Bericht FZKA 5670, Forschungszentrum Karlsruhe, Karlsruhe 1995, pp. 45 – 50.

55. K. Schubert: *"Entwicklung von Mikrostrukturapparaten für Anwendungen in der chemischen und thermischen Verfahrenstechnik,"* Bericht FZKA 6080, Forschungszentrum Karlsruhe, Karlsruhe 1998, pp. 53 – 60.
56. G. Wießmeier, D. Hönicke: "Microreaction Technology: Development of a Micro Channel Reactor and its Application in Heterogenously Catalyzed Hydrogenations," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 152 – 153.
57. G. Wießmeier: Dissertation, Universität Karlsruhe, 1996; Shaker-Verlag, Aachen 1997.
58. G. Wießmeier, K. Schubert, D. Hönicke: "Monolithic Microstructure Reactors Possessing Regular Mesopore Systems for the Successful Performance of Heterogeneously Catalyzed Reactions," in W. Ehrfeld (ed.): *Microreaction Technology, Proc. 1st Int. Conf. Microreaction Technology,* Springer-Verlag, Berlin 1997, pp. 23 – 25.
59. U. Hagendorf, M. Janicke, F. Schüth, K. Schubert, M. Fichtner: "A Pt/Al_2O_3 Coated Microstructured Reactor/Heat Exchanger for the Controlled H_2/O_2-Reaction in the Explosion Regime," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 81 – 87.
60. I. Robins, J. Shaw, B. Miller, C. Turner, M. Harper: "Solute Transfer by Liquid/Liquid Exchange without Mixing in Micro-contactor Devices," in W. Ehrfeld (ed.): *Microreaction Technology, Proc. of the 1st Int. Conf. on Microreaction Technology,* Springer-Verlag, Berlin 1997, 35 – 46.
61. J. Shaw et al.: "Characterisation of Micro-contactors for Solute Transfer between Immiscible Liquids and Development of Arrays for High Throughput," inR. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 267 – 271.
62. J. R. Burns, C. Ramshaw, A. J. Bull, P. Harston: "Development of a Microreactor for Chemical Production," in W. Ehrfeld (ed.): *Microreaction Technology, Proc. of the 1st Int. Conf. on Microreaction Technology,* Springer-Verlag, Berlin 1997, pp. 127 – 133.
63. J. R. Burns, C. Ramshaw, P. Harston: "Development of a Microreactor for Chemical Production," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 39 – 44.
64. W. E. Tegrotenhuis, R. J. Cameron, M. G. Butcher, P. M. Martin, R. S. Wegeng: "Microchannel Devices for Efficient Contacting of Liquids in Solvent Extraction," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 329 – 334.
65. K. Kadel, F. Götz, R. P. Peters: "Cross-Flow Filter Made using Deep Dry Etching," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 96 – 99.
66. K. Harre, A. Vuin, G. Wegner, K. Gebauer, T. Richter: "Membranmodul für Mikroreaktoren," in *Kurzfassungen DECHEMA-Jahrestagungen 1998, 26 – 28 May 1998, Wiesbaden,* DECHEMA e.V., Druckhaus K. Schmidt Wwe, Frankfurt/M. 1998, p. 269.
67. E. W. Becker: "The Separation Nozzle Process for Enrichment of Uranium-235," *Prog. Nucl. Energy* **1** (1977) 27 – 39.
68. A. L. Y. Tonkovich, J. L. Zilka, M. R. Powell, C. J. Call: "The Catalytic Partial Oxidation of Methane in a Microchannel Chemical Reactor," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 45 – 53.
69. T. Richter et al.: "Metallic Microreactors: Components and Integrated Systems," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 146 – 151.

70. K. F. Jensen et al.: "Reaction Engineering for Microreactor Systems," in *Proc. of 1st International Conference on Microreaction Technology, Feb. 23 – 25 1997, Frankfurt / M.* 1997, 2 – 9.
71. D. J. Quiram, I-Ming Hsing, A. J. Franz, R. Srinivasan, K. F. Jensen, M. A. Schmidt: "Characterization of Microchemical Systems using Simulations," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 205 – 211.
72. A. J. Franz et al.: "New Operating Regimes and Applications Feasible with Microreactors," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 33 – 38.
73. A. Wolf et al., "Mikroreaktorfertigung mittels Funkenerosion F&M, Feinwerktechnik, Mikrotechnik," *Meßtechnik* (1997) no. 6, 436 – 439.
74. M. Matlosz, C. Vallières: "Microsectioned Electrochemical Reactors for Selective Partial Oxidation," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 45 – 53.
75. M. Roth et al.: "Micro-structure Electrodes as Electronic Interface between Solid and Gas Phase: Electrically Steerable Catalysts for Chemical Reactions in the Gas Phase," in *Proc. of 1st International Conference on Microreaction Technology, Feb. 23 – 25 1997, Frankfurt / M.* 1997, pp. 62 – 69.
76. P. Woias et al.: "Silicon Microreactors with Photometric and Chemiluminescent On-column Detection," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 235 – 240.
77. P. Jacobs et al.: "Impedimetric Detection of Nucleic Acid Hybrids," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 223 – 229.
78. A. Manz: "The Secret behind Electrophoresis Microstructure Design," in Proceedings of the 2nd International Symposium on Miniaturized Total Analysis Systems µTAS96, Basel, *Analytical Methods & Instrumentation AMI,* 1996, special issue, 28 – 30.
79. K. Schmidt, R. Ehricht, T. Ellinger, J. S. McCaskill: "A Microflow Reactor with Components for Mixing, Separation and Detection for Biochemical Experiments," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 125 – 126.
80. J. M. Köhler, U. Dillner, A. Mokansky, S. Poser, T. Schulz: "A Microflow Reactor with Components for Mixing, Separation and Detection for Biochemical Experiments," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference Preprints, March 9 – 12 1998,* New Orleans 1998, pp. 241 – 247.
81. O. T. Guenat, P. Arquint, W. E. Morf, B. H. van der Schoot, N. F. de Rooij: "Coulometric Nanotitrators with Potentiometric Endpoint Detection," in *Proc. of 1st International Conference on Microreaction Technology, Feb. 23 – 25 1997, Frankfurt / M.* 1997, 340 – 347.
82. M. A. Northrup, B. Benett, D. Hadley, P. Stratton, P. Landre: "Advantages Afforded by Miniaturization and Integration of DNA Analysis Instrumentation," in *Proc. of 1st International Conference on Microreaction Technology, Feb. 23 – 25 1997, Frankfurt / M.* 1997, pp. 278 – 288.
83. A. van den Berg, P. Bergveld: "Development of µTAS Concepts at the MESA Research Institute," in Proceedings of the 2nd International Symposium on Miniaturized Total Analysis Systems µTAS96, Basel, *Analytical Methods & Instrumentation AMI,* 1996, special issue, 9 – 15.
84. J. M. Ramsey, S. C. Jacobson, R. S. Foote: "Microfabricated Devices for Performing Chemical and Biochemical Analysis," in *Proc. of 1st International Conference on Microreaction Technology, Feb. 23 – 25 1997, Frankfurt / M.* 1997, 204 -218.
85. N. Schwesinger, O. Marufke, F. Qiao, R. Devant, H. Wurziger: "A Full Wafer Silicon Microreactor for Combinatorial Chemistry," in W. Ehrfeld, I. H. Rinard, R. S. Wegeng (eds.): *Proc. of 2nd International Conference on Microreaction Technology, Topical Conference*

Preprints, March 9 – 12 1998, New Orleans 1998, pp. 124.
86. L. Dubois: "From Mesoscopic Machines to Microchemical Synthesis: DARPA's Interest in Small-Scale Systems," presentation given at *Proc. of 2nd International Conference on Microreaction Technology, March 9 – 12 1998,* New Orleans 1998.

Microscopic Stains → **Stains, Microscopic**

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Sonochemistry

Standard Article

Urs A. Peuker[1], Ulrich Hoffmann[1], Ulrich Wietelmann[2], Stefan Bandelin[3], Rainer Jung[3]

[1]Institut für Chemische Verfahrenstechnik, Clausthal University of Technology, Germany

[2]Chemetall GmbH, Frankfurt, Germany

[3]Bandelin electronic GmbH & Co. KG, Berlin, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.b24_b01.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Fundamentals of Ultrasound
2.1. Ultrasound in Liquids
2.2. Cavitation
2.2.1. Bubble Behavior in Acoustic Fields
2.2.2. Cavitation Thresholds
2.2.3. Parameters Influencing Cavitation
2.3. Nonlinear Effects
2.3.1. Bubble Collapse near Boundaries
2.3.2. Finite Amplitude Waves and Shock Waves
2.3.3. Streaming
2.3.4. Radiation Forces
2.3.5. Bjerknes Forces
2.3.6. Forces on Small Particles
2.4. Sonochemical Effects
2.4.1. Hot-Spot Theory
2.4.2. Shock Wave Theory
2.4.3. Supercritical Water Theory
2.4.4. Charge Theory
2.4.5. Promotion of Single Electron Transfer
2.4.6. Cleaning
2.4.7. Mechanical Activation
2.4.8. Enhanced Mass and Heat Transfer
3. Heterogeneous Systems
3.1. Solid - Fluid Reactions
3.1.1. Noncatalytic Reactions
3.1.2. Catalytic Reactions
3.1.3. Precipitation and Crystallization
3.1.4. Atomization
3.1.5. Disaggregation – Comminution
3.1.6. Cell Disruption
3.2. Ultrasonic Extraction
3.3. Liquid - Liquid Reactions
3.3.1. Emulsification
3.4. Phase-Transfer Catalysis

3.5. Sonoelectrochemistry
4. Homogeneous Systems
4.1. Aqueous Liquids
4.2. Nonaqueous Liquids
4.3. Polymers and Macromolecules
4.4. Sonoluminescence
5. Equipment
5.1. Transducer Technology
5.2. Reactors for Sonochemical Reactions
5.2.1. Laboratory Equipment
5.2.2. Large-Scale Equipment
5.3. Mapping of Sonochemical Reactors
5.3.1. Hydrophone
5.3.2. Thermocouples
5.3.3. Optical Methods – Bubble Dynamics
5.3.4. The Weissler Reaction

Applications of ultrasound in processing and synthesis are widespread and offer unusual and beneficial operating conditions. Two major principles are used in sonochemistry or sonoprocessing: a) mechanical effects for mixing and disintegration and b) high energy processes for radical reactions. The fundamentals in ultrasound and the parameters influencing cavitation are explained.

In heterogenous systems, ultrasound is used in noncatalytic and catalytic reactions, biological applications, crystallization, emulsification, atomization, extraction, electroplating, phase-transfer catalysis. In homogeneous systems, applications of ultrasound include the preparation and degradation of polymers and macromolecules.

Mapping of sonochemical reactors can be achieved with several measuring systems and provides an essential knowledge about intensity distribution of the pressure field within the reactor volume.

[Top of Page]

1. Introduction

Applications of ultrasound in processing and synthesis are widespread and offer unusual and beneficial operating conditions. Two major principles are used in sonochemistry or sonoprocessing: a) mechanical effects for mixing and disintegration and b) high energy processes for radical reactions. Most effects occur in liquids, where sound pressures are able to disrupt the continuum. This generates oscillating bubbles, a process called cavitation. Cavitation is observed in shock tubes, pumps and other hydrodynamic devices. Due to the high frequencies in sonochemistry of between 16 kHz and several megahertz, the dynamics of oscillating bubbles create drastic conditions. Temperatures of several thousand Kelvin, extreme heating or cooling rates of 10^5 K/s, and pressures of up to several hundred megapascal are observed in a transient cavities while the bulk conditions in the liquid remain at ambient temperature and pressure. Radiation forces create intense micro and macromixing with high shear forces, which are used in emulsification, homogenization, and fragmentation processes. Asymmetrical bubble oscillations in the vicinity of solid particles lead to liquid microjets and shock waves, which are used in cleaning, dispersion, activation, and fragmentation of solid materials. Besides the main industrial applications in cleaning, emulsification, and welding, other uses of ultrasound, such as solids processing, atomization, crystallization, environmental protection and separation are emerging (Table 1). Some synthetic methods, for example the Barbier and Grignard reactions, have found industrial utilization. The growing interest in ultrasound application arises from the development of reliable ultrasound devices in the last years and the need for intensified and environmentally friendly processes with small safe reactor units, better energy yield, and enhanced chemical yield.

Table 1. Applications of ultrasound in processing

Field	Applications
Heterogeneous reactions	activation of solids in reactions, fragmentation, dispersion, agglomeration, preparation of catalysts, hydrogenation at room temperature
Homogeneous reactions	generation of radicals, oxidation of wastewater compounds, nanoparticle synthesis
Polymer processing	controlled degradation, initiation of polymerization, copolymerization, mixing in suspension polymerization
Electrochemistry	avoidance of mass transport limitations, electrode fouling, change in reaction pathways
Cleaning	cleaning of items in aqueous and nonaqueous media
Mixing	dispersion of solids, particle agglomerates, and pigments; mixing of immiscible liquids
Crystallization	initiation of seeding, control of particle size distribution, control of material

	properties
Extraction	fast and mild extraction of essential oils or microbiological material
Welding	welding of thermoplastics and metals
Cutting	cutting of ceramics and metals
Separation	improved filtration, agglomeration, and separation of particles

[Top of Page]

2. Fundamentals of Ultrasound

The term ultrasound describes sound waves with a frequency range from 16 kHz up to several megahertz [2]. Vibrational motions are transmitted by oscillating devices into a fluid and cause pressure waves. This varying sound pressure is superimposed on the static pressure of the liquid. Fluids are generally capable to transduce longitudinal waves, whereas solids also can show transversal motion. Above a critical sound pressure, liquids are disrupted by the applied sound pressure and so-called cavities are created. Most sonochemical effects are secondary effects generated by oscillating bubbles.

2.1. Ultrasound in Liquids

Sound waves in liquids are density and pressure waves in which particle oscillations occur in the direction of the wave (longitudinal waves). The displacement around the rest position causes compression and rarefaction disturbances which are transmitted to the connected layers (Fig. 1). The time-dependent particle displacement ξ the particle velocity v and the sound pressure p_a for plane waves can be described by:

$$\xi = \hat{\xi} \sin(2\pi f t) \tag{1}$$

$$v = \frac{d\xi}{dt} = \hat{v} \cos(2\pi f t) \tag{2}$$

$$p_a = \hat{p}_a \cos(2\pi f t) \tag{3}$$

with the maximum values of particle displacement, particle velocity and sound pressure indicated by circumflexes. In plane waves, the particle velocity v and the sound pressure p_a are related by the specific impedance of the liquid

$$\varrho_L c_L = \frac{p_a}{v} \tag{4}$$

which is the product of the liquid density ϱ_L and velocity of sound c_L.

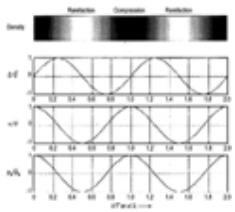

Figure 1. Propagation of a traveling wave with density, normalized particle displacement, particle velocity and sound pressure as a function of normalized time or coordinate

The sound wave transports energy by kinetic energy of the oscillating particles. If we assume an energy density E given by

$$E = 0.5\, \varrho_L v^2 \tag{5}$$

which passes through a cross-sectional area S with the velocity of sound c we derive the so-called ultrasound intensity I as the energy flux per unit area:

$$I = 0.5\, \varrho_L c_L \hat{v}^2 = 0.5 \frac{\hat{p}_a^2}{\varrho_L c_L} \tag{6}$$

The determination of local intensities is quite difficult, and ultrasound devices are mostly characterized by calorimetric measurements. Sound pressure measurements and methods for determining the local intensity with coated thermocouples or by chemical means are still under development [1].

2.2. Cavitation

Sufficiently high sound pressures in liquids create voids or gas- and vapor-filled bubbles. Any liquid has a theoretical tensile strength which characterizes the minimum pressure for disruption. Due to the presence of nuclei such as dissolved gases, solid impurities, and rough walls, cavitation occurs at far lower sound pressures than are theoretically necessary. In nearly any liquid, initial nuclei are present which show a distinct size distribution and grow under a certain sound pressure. Bubble growth, multiplication, and disappearance in a sound field is still very complex phenomenon.

2.2.1. *Bubble Behavior in Acoustic Fields*

Several types of bubbles are present in cavitating liquids. Empty cavities (true cavitation), gas-filled cavities, vapor-filled cavities or mixtures of gas and vapor can be observed, depending on the applied sound pressure, static pressure,

temperature, and the nature of the used liquid. Some of these bubbles disappear because of dissolution of their contents under the sound pressure, some of them oscillate stably over several periods and others collapse violently, followed by the creation of smaller bubbles. The main task in sonochemical applications is the choice of the proper bubble behavior for the desired effect. Bubble motions in a sinusoidal sound pressure field can be described, for example, by the Rayleigh – Plesset equation

$$R\ddot{R}+\frac{3}{2}\dot{R}^2=\frac{1}{\varrho_L}\left\{\left(p_\infty+\frac{2\sigma_L}{R_0}-p_\nu\right)\left(\frac{R_0}{R}\right)^{3\gamma}+p_\nu-\frac{2\sigma_L}{R}-\frac{4\eta_L\dot{R}}{R}-p_\infty-p_a\right\} \quad (7)$$

which includes the surface tension σ_L, the liquid viscosity η_L, the polytropic exponent γ of the gas inside the bubble, the vapor pressure p_V and the initial bubble radius R_0 at rest. The solutions of this equation lead to nonlinear bubble wall motions (Fig. 2). A special case are resonant bubbles, which show damped oscillations. The resonant radius R_r of such a bubble can be described by Minnaert's formula

$$\varrho_L\omega_r^2R_r^2=3\gamma\left(p_\infty+\frac{2\sigma_L}{R_r}\right)-\frac{2\sigma_L}{R_r} \quad (8)$$

with the circular frequency $\omega_r = 2\pi f$.

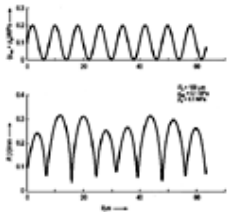

Figure 2. Total pressure (top) and solution of the Rayleigh – Plesset equation for bubble radius (bottom) of a stable bubble as a function of time for air saturated water at 20 kHz with an initial bubble radius R_0 of 0.1 mm, a static pressure of 0.1 MPa and a sound pressure of 0.1 MPa

2.2.2. *Cavitation Thresholds*

A more illustrative description of the bubble motion in an acoustic field can be introduced by the concept of cavitation thresholds for different types of bubble behavior [10]. Depending on the nature of the motion four basic types of cavitating voids are distinguished (Fig. 3):

- Stable cavitation
- Rectified diffusion
- Dissolving bubbles
- Transient cavitation

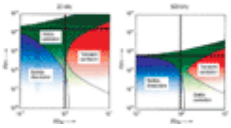

Figure 3. Simplified cavitation thresholds and regions for air saturated water at a static pressure of 0.1 MPa at frequencies of 20 and 500 kHz

In the absence of an acoustic field, a gas bubble in a liquid will slowly dissolve owing to the excess internal pressure required to withstand the surface tension pressure $2\sigma_L/R_0$. By applying a changing pressure, the dissolution time can be increased. Very small bubbles will still dissolve, but some bubbles larger than a certain threshold radius can be stabilized in the sound field. These stable oscillating bubbles survive several acoustic cycles. In a process, known as rectified diffusion, such bubbles collect gas from the liquid and grow in size. In the rarefaction cycle, the bubble radius increases and volatile gases or vapor can enter the bubble across the interface. In the following compression cycle, the bubble shrinks, and the diffusion of gases into the liquid is hindered by the smaller exchange surface. This effect is aided by the changing thickness of the liquid boundary layer, which generates high concentration gradients at large radii and small gradients in the compression phase. Rectified diffusion is responsible for the degassing of liquids under ultrasound, even at low sound pressure of 0.01 MPa. Such a growing bubble may reach the resonance radius, at which very strong oscillations cause surface instabilities and the generation of smaller bubbles. Surface oscillations are responsible for very high microstreaming in the vicinity of stable bubbles, which accelerates heat and mass transport. The behavior of some bubbles can change dramatically at a certain radius. The radius grows rapidly to more than twice the size in half an acoustic period. This is followed by a sudden collapse after which the bubble disappears. These bubbles are known as transient bubbles (Fig. 4) and cause most of the sonochemical effects.

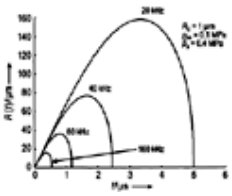

Figure 4. Bubble radius as a function of time for a transient cavity with an initial radius of R_0 = 1 µm, a static pressure of 0.1 MPa, and a driving sound pressure amplitude of 0.4 MPa in air-saturated water. The maximum bubble radius decreases with increasing frequency

Transient bubbles have a very small collapse time which was first deduced by RAYLEIGH. An approximation for the collapse time in acoustic sound fields is

$$\tau_B = 0.915 \cdot R_{max}\left(\frac{\varrho_L}{p_m}\right)^{\frac{1}{2}}\left(1+\frac{p_v}{p_m}\right) \tag{9}$$

where R_{max} is the maximum bubble radius before collapse, p_m the mean driving pressure (the sum of hydrostatic and sound pressure), and p_V the vapor or gas pressure inside the bubble at maximum expansion. Transient bubble generation is subject to two restrictions: First, a transient bubble must undergo extensive growth, for which its radius must exceed a certain threshold at which the forces acting on it are higher than the surface tension forces. Secondly, having reached this state, sufficient energy must be concentrated in a very short time, and this requires critical values of the radius before (R_{max}) and after collapse (R_{min}). Only in this case, are the energy dissipation (heat conduction, viscosity losses) smaller than the work done by the spherical convergence of the surrounding liquid. These two criteria are described by the so-called Blake threshold radius $R_{B,0}$ and the Flynn dynamic radius $R_{F,0}$. Bubbles having initial radii R_0 smaller than the Blake threshold and greater than the Flynn dynamic radius will undergo transient collapse. APFEL used these criteria and derived an expression for the maximum bubble radius before collapse.

$$R_{max} = \frac{4}{3\omega}(\hat{p}_a - p_\infty)\sqrt{\frac{2}{\varrho_L \hat{p}_a}} \left(1+\frac{2}{3p_\infty}(\hat{p}_a + p_\infty)\right)^{\frac{1}{3}} \tag{10}$$

Assuming isothermal expansion to maximum bubble radius R_{max} and subsequent adiabatic collapse, the pressures and temperatures generated in a gas-filled transient cavity can be derived as

$$T_{max} = T_0(\gamma - 1)\left(\frac{R_{max}}{R_0}\right)^3 \tag{11a}$$

$$p_{max} = p_{g,m}\left(\frac{p_\infty(\gamma-1)}{p_{g,m}}\right)^{\frac{\gamma}{\gamma-1}} \tag{11b}$$

where $p_{g,m} = p_\infty(R_0/R_{max})^3$ is the pressure in the bubble at maximum radius.

2.2.3. *Parameters Influencing Cavitation*

Influencing parameters in sonochemical applications are listed in Table 2.

Table 2. Important influencing parameters in sonochemical applications

Parameter	Property	Effect
Dissolved gas	polytropic ratio	cavitation intensity
	thermal conductivity	primary sonochemistry
	solubility	cavitation intensity
Vapor pressure	content of bubbles	cavitation intensity
Liquid viscosity	dissipative effects	cavitation intensity
Temperature	liquid properties	cavitation intensity
	thermal activation	secondary reaction rate
Static pressure	thresholds	cavitation intensity and
	gas solubility	extent of cavitation zone
Frequency	thresholds	cavitation intensity
	wavelength	extent of cavitation zone
Intensity	thresholds	cavitation intensity
	attenuation	extent of cavitation zone
	maximum bubble radius	cavitation intensity

Dissolved Gases are present in most applications of ultrasound. The type and amount of these gases in the treated liquid are a very important factor for the sonochemical effects [11]. Highly gas saturated liquids have large numbers of cavitation nuclei and low cavitation thresholds. Degassed liquids need higher sound pressures for cavitation, but the sonochemical effects are in most cases more intense. The bubbles can collect gas from the surrounding liquid, and this rectified diffusion creates large bubbles with high gas contents. On collapsing, these bubbles create only moderate temperatures and pressures. As a general rule, degassing a liquid is always favorable in sonochemical applications. Beside the gas content of a cavitating bubble, the used gas is important (Table 3). Gases with high polytropic exponents and small thermal conductivity show the best effects under sonication. Monoatomic gases like argon and xenon exhibit good cavitation intensities, whereas diatomic gases such as nitrogen and oxygen tend to decrease the observed effects. Gases with high thermal conductivity

such as helium decrease the intensity of cavitational collapse to nearly zero.

Table 3. Properties of different gases and their influence on H_2O_2 yield and sonoluminescence intensity

Gas	Specific heat ratio γ	Thermal conductivity, $mWm^{-1}K^{-1}$	Solubility in water at 20 °C, mL/100 mL H_2O	Relative H_2O_2 yield [11]	Relative sonoluminescence intensity [5]
Nitrogen	1.40	25.83	2.33	2.5	45
Oxygen	1.40	26.46	4.89	13.5	35
Helium	1.65	149.68	0.94	1	1
Neon	1.67	48.02	2.6	7	18
Argon	1.67	17.75	5.6	21.5	54
Krypton	1.67	9.46	11.0	24.0	180
Xenon	1.67	5.19	24.1	27.5	540

Vapor Pressure. The vapor pressure of liquids can cushion the bubble collapse like a high gas content. Vapor in a transient bubble can be condensed in the compression cycle and lead to higher cavitation intensities than gas filled bubbles. Experiments with different solvents show that small vapor pressures are necessary for a sufficiently high cavitation intensity. Higher vapor pressures, especially near the boiling point of the liquid, can dampen the cavitation efficiency to nearly zero. If a substrate is subject to treatment within the collapsing bubbles, then a certain amount of its molecules must be present in the bubbles and exert an at least measurable vapor pressure. The existence of molecules inside the bubble can easily be proved by means of molecules which exist as ionic or molecular species at different pH values. Ionic species do not enter the bubbles, and high temperature pyrolysis products can therefore not be created.

Viscosity. High viscosity of a sonicated liquid increases the cavitation threshold markedly. Viscous liquids generate bubbles only at high sound pressures. Bubble motion is damped by the dissipative effect of the viscosity and the smaller maximum bubble radii, and the lower inward wall velocities terminate most sonochemical effects.

Temperature. The effect of temperature on sonochemical processes is explained by its influence on viscosity, gas solubility, vapor pressure and surface tension. Any chemical reaction will be influenced by the temperature-dependent reaction rate coefficient. In most cases, cavitation is more pronounced at lower temperatures due to the lower vapor pressure of the liquid. This lowers the total gas content of the collapsing bubbles and causes higher cavitation intensities. This paradoxical temperature dependence is highly pronounced for sonochemical reactions involving radical generation in transient bubbles. Lesser effects are generally observed in heterogeneous systems, where intense micromixing by oscillating bubbles favors mass and heat transfer.

Static Pressure. The static pressure in a sound field alters the thresholds for rectified diffusion, transient bubbles and other characteristics. A high static pressure can prevent the generation of bubbles by ultrasound or create a different size distribution. For a given amount of energy, smaller bubbles with a higher energy content are created, which show very strong erosion activities in heterogeneous systems.

Frequency. The effect of frequency on sonochemical reactions is still an active field of research. Some measurements indicate that different frequency ranges are needed for special reaction types. Low frequencies in the range from 16 to 100 kHz, known as power ultrasound, are mainly active in heterogeneous systems with micromixing, cleaning, mechanical action on suspended solids and intense bubble motion. High frequencies favor high temperatures and pressures in the cavitation bubbles, thus creating a large number of radicals. The sonolysis of water seems to be most pronounced at frequencies between 300 and 500 kHz. Frequency is a very important parameter in defining whether a bubble of a given size is transient or stable. The smaller bubbles present in a high frequency sound field have a lower energy content than the larger bubbles generated at low frequencies. Beneficial for sonochemical effects is the higher number of cavitational events in high-frequency applications. In nearly all sonochemical processes, an optimum frequency can be found experimentally.

Ultrasound Intensity. The effect of intensity or power input to the ultrasonic device is complicated. Higher intensities provoke a larger amplitude at the vibrating surface in contact with the liquid. At low intensities, a linear dependency between generated sound pressure and amplitude is observed. Raising this intensity above the cavitation threshold of the liquid causes oscillating bubbles, and under certain circumstances the contact between radiating surface and liquid is lost. The motion of the transducer and the liquid are out of phase, an effect known as decoupling. A second reason for the nonlinear relationship between intensity and sonochemical effects is the creation of cavitation zones. Higher intensities create more and larger bubbles, which may coalesce and lead to less transient events. Energy efficiency and sonochemical effects are therefore sometimes smaller at higher intensities, and often the optimum intensity for a specific sonochemical effect must be determined experimentally.

Pulsed Ultrasound. Pulsed ultrasound is characterized by phases with and without ultrasound. In the ultrasound cycle, bubbles are created, grow by rectified diffusion, and may eventually collapse. The silent cycle can be used to remove larger bubbles by Archimedian forces. Smaller bubbles begin to dissolve, and the overall gas content decreases, giving optimum conditions for the next sonication cycle. Many sonochemical processes benefit from pulsed operation and show better performance than with continuous operation.

2.3. Nonlinear Effects

2.3.1. *Bubble Collapse near Boundaries*

Undisturbed bubbles in liquids show spherical oscillations in a sound field [12]. Any extended disturbances such as boundaries, suspended solids, or reactor walls in the vicinity of the transient bubble prevent spherical collapse. Asymmetric bubble wall motion leads to the formation of an involution, which is directed towards the disturbance. In the late stages of the collapse of a transient bubble, a liquid microjet breaks through the remote bubble wall and impinges on the boundary. Liquid microjets can reach velocities in the range of several hundred meters per second and are responsible for cavitation erosion on solids. The pressures generated on solid surfaces can melt soft metals, create pitting, or disrupt inorganic layers on metals. A liquid jet with a diameter of roughly one tenth of the collapsing bubbles diameter creates a water hammer p_{WH} pressure of

$$p_{WH} = \frac{(\varrho_L c_s \varrho_T c_T)\, v_J}{\varrho_L c_s + \varrho_T c_T} \tag{12}$$

where c_S is the shock speed ($c_S \approx c_L + 2v_J$), v_J the jet speed, c_T the speed of sound and ϱ_T the density of the target. This pressure is maintained in the center of the jet for a time $t_{col} = R_{max}/10c_S$ followed by a Bernoulli pressure of $\varrho_L v_J^2/2$. Most heterogeneous reactions make use of these mechanical effects in cavitating sound fields. Even more intense effects are observed when clouds of bubbles collapse near solid boundaries. The concerted breakdown of such a hemispherical transient bubble cloud generates shock wave pressures and jetting with intensities some orders of magnitude higher than those of single bubbles [13].

2.3.2. *Finite Amplitude Waves and Shock Waves*

Most applications in ultrasound deal with rather small particle velocities v. This can be described by the acoustic Mach number $M = v/c_L$. A 20 kHz high-intensity step horn with an amplitude of 100 µm creates a Mach number of about 0.01. For low Mach numbers, the acoustic approximation of incompressible liquids is applicable. In certain cases, for example focused sound fields, much higher Mach numbers are present and lead to distortion of the sinusoidal wave form. Density and sound pressure no longer satisfy a simple linear equation with a fixed speed of sound c. A pressure-dependent speed of sound is observed (material nonlinearity). Convective terms in the equation of motion cannot be neglected at higher Mach numbers and lead to a more complicated situation (convective nonlinearity) in which the real speed of the wave front consists of two parts, the speed of sound c and the particle velocity v. The material nonlinearity and the convective nonlinearity tend to increase the propagation velocities of zones with high pressures. This leads to saw-tooth waveforms which develop a shock wave. Consisting of higher harmonics of the initial sound frequency, these saw-tooth waves are strongly attenuated and quickly decay into low-amplitude sinusoidal waves with the initial frequency. Shock waves are also formed in transient bubble collapse. If the bubble wall reaches high velocities, a shock wave is emitted into the surrounding liquid. The Mach number in this case is $M = (dR/dt)/c_L$ and reaches values higher than unity. The shock wave is generated some distance from the bubble. Hickling and Plesset calculated, that above a pressure of 100 MPa inside the bubble, a shock wave starts at a distance of $r/R_{min} = 5 – 6$ where R_{min} is the minimum bubble radius at collapse [14].

2.3.3. *Streaming*

Acoustic Streaming. As an acoustic wave travels through a medium, it can be absorbed. Because of the absorption of momentum in the direction of the sound field, flow is initiated in this direction. Local variations in intensity and energy absorption lead to local streaming velocities in the order of several centimeters per second. Finite amplitude waves create acoustic streaming due high absorption of higher harmonic components.

Acoustic Microstreaming. Small obstacles in a sound field create circulation by friction between their boundaries and the vibrating liquid particles. This acoustic microstreaming boundary layer has a thickness δ_{MS} of

$$\delta_{MS} = \sqrt{\frac{2\eta_{L,S}}{\varrho_L \omega}} \tag{13}$$

where $\eta_{L,S}$ is the shear viscosity of the liquid. Microstreaming enhances mass and heat transfer and leads to shear forces on the obstacles.

2.3.4. *Radiation Forces*

Objects in an acoustic sound field experience a force whose magnitude and direction depends on the intensity of the source if the dimensions of the target are greater than the wavelength. Two types of radiation forces can be distinguished, depending on the boundary conditions. The Rayleigh radiation force acts on a target of area S in a closed vessel with energy density E:

$$F = 0.5\,(1+\gamma)\,ES \tag{14}$$

The case of open vessels (free boundary condition) is treated by the Langevin radiation force

$$F = ES \tag{15}$$

which represents most experimental conditions. The Langevin radiation force is utilized to measure the intensity of sound fields.

2.3.5. *Bjerknes Forces*

Stable oscillating bubbles experience translational radiation forces in a traveling-wave field. In standing-wave fields, these so-called primary Bjerknes forces act on oscillating bubbles and are due to a time-average force resulting from sound field pressure gradients and bubble oscillation. Bubbles below resonance size are attracted to pressure antinodes, whereas

bubbles above resonance size travel down pressure gradients towards pressure nodes (Fig. 5) [33].

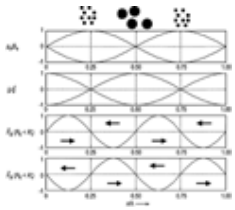

Figure 5. Normalized sound pressure, particle displacement and primary Bjerknes forces acting on bubbles smaller and greater than the resonance radius R_0. Small bubbles collect in pressure antinodes (displacement nodes), whereas large bubbles travel toward pressure nodes

Secondary Bjerknes forces arise when two oscillating bubbles are present in a pressure field. Attractive forces between bubbles with in-phase pulsation cause coalescence. Bubbles oscillating out of phase are repelled. The bubble oscillation is in phase when both bubbles are smaller or larger than resonance size and attractive forces dominate. If one bubble is smaller and one larger than resonance size, they oscillate in out of phase and repel one another.

Primary and secondary Bjerknes forces lead to structures known as cavitation streamers. A large bubble above resonance size oscillating with surface instabilities and located in a pressure node generates microbubbles below resonance size, which are repelled and forced towards the pressure antinode of a standing-wave field, where attractive forces cause clouds of oscillating microbubbles. The lines of microbubbles are often called microstreamers.

2.3.6. *Forces on Small Particles*

Small particles with radii smaller than the cavitation bubbles are accelerated in the velocity and pressure gradients around oscillating bubbles. These forces are even greater when particles are subject to shock waves. Particle velocities of up to 500 km/h can be reached. On collision of two such particles, oxide layers are removed and metals are melted together. Such treated metals are used in catalysis, where activity enhancements by a factor of up to 10^6 are observed.

Like the Bjerknes forces, solid particles whose acoustic properties differ from those of the liquid are subject to primary and secondary forces. Primary forces drive particles into pressure nodes, where secondary forces between particles are responsible for further aggregation [15].

2.4. Sonochemical Effects

The effects of ultrasound on chemical reactions are in most cases secondary effects caused by cavitation and can be divided into five main groups:

1. Reactions inside the cavitation bubble at high temperatures and pressures
2. Reactions at the gas – liquid interface of the bubbles (secondary reactions of products formed within the bubbles, high-temperature reactions of nonvolatile liquids)
3. Reactions due to high pressures in the surrounding liquid
4. Effects caused by nonlinear bubble collapse near boundaries
5. Enhanced mass and heat transfer by macro- and micromixing

2.4.1. *Hot-Spot Theory*

Very high temperatures and pressures are generated in a cavity on collapse. Noltking and Neppiras calculated temperatures of up to 10^4 K in a transient cavity. This would lead to black-body thermal radiation which can be observed spectroscopically. Temperatures of 5000 K in the bubble, 2000 K in the liquid boundary layer and pressures of 50 MPa were found in experiments with metal carbonyles [5]. UV/Vis spectroscopy of the cavitating liquid showed good agreement with the thermal spectra of carbon and the metals at the same temperature. Thus, by creating extreme temperatures in transient bubbles and the surrounding liquid boundaries, high-temperature reactions can be initiated in a liquid that remains at ambient temperature. These calculations are confirmed by the presence of pyrolysis products of irradiated liquids. These include the homolytic cleavage of water, the cleavage of carbon-halogen bonds, and the production of radical products from organic liquids like alkynes and alkenes. The extreme heating and cooling rates of about 10^5 K/s resemble those in the freezing of molten metals when poured onto a plate with a temperature of – 270 °C. Under the extreme conditions of a transient bubble, even plasma reactions or fusion were proposed. Theoretically, conditions similar to those on the surface of the sun are possible.

2.4.2. *Shock Wave Theory*

The compression of transient bubbles leads to an increasing bubble wall velocity which may eventually reach the speed of sound of the liquid. In true transient cavitation, the bubble vanishes after collapse, creating a shock wave in the liquid. Particles and macromolecules are accelerated in the steep pressure gradient and are shock fragmented. High speed particles collide and undergo mechanical damage.

2.4.3. *Supercritical Water Theory*

High temperatures and pressures in aqueous sonochemistry may lead to conditions under which supercritical water is likely. Supercritical water is known to have a strong solvent action towards organic compounds and extreme chemical activity. Sonochemical effects are possible inside the supercritical water layer surrounding a transient bubble. At the time being, no direct evidence for the generation of supercritical water in ultrasound fields has been found experimentally.

2.4.4. *Charge Theory*

Russian workers questioned the existence of true cavitation and postulated a somewhat different sonochemical effect. Oscillating bubbles undergo very rapid size changes. Friction forces at the gas-liquid bubble boundary can create charged

species which can lead to secondary chemical reactions in the bulk liquid.

2.4.5. *Promotion of Single Electron Transfer*

Classical ionic bimolecular nucleophilic substitution reactions (S_N2) have been interpreted as a radical mechanism with radical formation by transfer of a single electron according to

$$RX + Y^- \rightarrow RX^{\cdot -} + Y \qquad (16)$$

Ultrasound has a accelerating effect on such single-electron transfer reactions (SETs) and can alter the reaction pathways when ionic S_N2 reactions and SET reactions occur simultaneously and lead to different products.

2.4.6. *Cleaning*

Microstreaming, shock waves, and liquid microjets in the vicinity of solid surfaces lead to very efficient cleaning. This effect has been used in industry for more than forty years. Insoluble layers of inorganic salts, polymers, or liquids can be removed by ultrasonic cleaning effect. In heterogeneous systems such a clean reactive surface leads to improved dissolution rates of metals in acids and enhanced reaction rates. Chemical reactions giving insoluble products are freed from these mass transport limiting layers and react rapidly.

2.4.7. *Mechanical Activation*

Several mechanisms for the activation of solids in chemical reactions are induced by the mechanical effects of ultrasound and transient cavities. High intensity ultrasound is able not only to remove passivating layers on solids, but to break solids like salts and metals as well (Fig. 6).

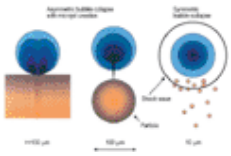

Figure 6. Influence of transient bubbles on particles of different size in a 20 kHz sound field. Extended surfaces lead to asymmetrical collapse with formation of microjets. Medium-sized particles to less asymmetrical collapse. Small particles are accelerated by bubble oscillation and in the shock wave

Surface Renewal. Impacting liquid microjets or shock waves create the pitting on solid surfaces [17]. These indentations are free of unreactive layers and react readily with dissolved reactants. Mechanical energy is introduced to the lattice of the solid and leads to slightly excited active centers. In general, such activated solids exhibit higher reaction rates than the untreated solids.

Particle Disruption. Brittle solids can be cleaved by transient cavities. Like in milling operations, this leads to higher specific surfaces, high energy content at the disrupted site, and reactive fresh surfaces. Because of the tiny bubble size, very small particles in the micrometer range can be produced.

Particle Agglomeration. Acceleration of small particles in the shock wave field of a transient bubble can lead to particle agglomeration. Highly accelerated particles collide and melt together on contact. Contact temperatures of nearly 3000 K were measured for very small particles. Particle agglomerates in the vicinity of transient bubbles are fragmented and yield higher specific surfaces [16]. Primary and secondary forces on suspended solids can increase the amount of aggregation in pressure antinodes of a standing wave field. These effects are used, for example, in separation processes.

2.4.8. *Enhanced Mass and Heat Transfer*

Many heterogeneous reactions are accelerated by the enhanced micromixing properties of cavitating sound fields. Oscillating and transient bubbles create intense microstreaming in the vicinity of suspended solids. Macromixing is induced by acoustic streaming and the oscillation of bubbles in the sound field. In most cases, a locally different mass transport coefficient is observed. Tenfold increases in mass-transfer coefficients compared with silent reactions were measured [18].

[Top of Page]

3. Heterogeneous Systems

3.1. Solid – Fluid Reactions

In heterogeneous solid – fluid reactions, asymmetrical bubble collapse creates shock waves and liquid microjets. These are responsible for particle fragmentation, pitting, and cleaning of the solid (see ⟶ Metals, Surface Treatment – Cleaning). Particles that are larger than the transient cavity must be present to induce asymmetrical collapse. Powdered solids which are smaller than the transient bubble are accelerated in the shock waves surrounding a cavity. High impact speeds lead to cleaning, shock fragmentation and mechanical activation. Metal particles in the micrometer range are melted together. Only tungsten (*mp* 3410 °C) resists aggregation by welding [19]. In a 20 kHz sound field for example, extended particles have to be roughly 200 µm in size to perturb the radial collapse. In this case, loss of material and some surface hardening of metals is observed. Surface hardening is due to an increasing lattice deformation with higher defect density on the surface of the solid. This high defect density also leads to a better chemical reactivity for specific reactions and facilitates syntheses by solid – fluid reactions, especially in organometallic chemistry.

3.1.1. *Noncatalytic Reactions*

Alkali metals are often used in synthesis and are very suitable for ultrasonic processes [8]. Passivating layers on the pure metals are removed, and the induction times of reactions are shortened dramatically. Aromatic radical anions like naphthalene – lithium, anthracene – sodium or biphenyl – sodium are easily formed under ultrasound [9] in 45 min, compared

to 48 h without ultrasound. Very fine dispersions of sodium and potassium are possible in solvents with low vapor pressures such as toluene or xylene. Dispersed colloidal metals are used in Dieckmann condensations and Wurtz coupling.

The Barbier reaction needs under classical conditions dried solvents like THF or ether, an organic chloride, and lithium wire or powder [20]. With ultrasound, moist solvents can be used. A reduction of induction time, high reaction rates and high selectivity are observed because sonication is performed under mild conditions below the boiling point of the solvent. In most cases, no difference in the total yield was observed. One of the advantages of the Barbier reaction is the possibility to produce highly unstable intermediates in situ which will react with a second reagent:

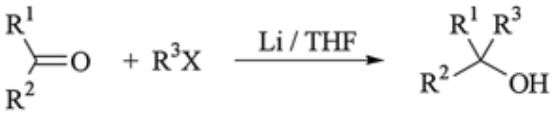

(17)

The Bouveault Reactions of alkyl, cycloalkyl, and aryl bromides with Li and DMF in ether solvents to form aldehydes is drastically enhanced by ultrasound, and high selectivity is observed [8]:

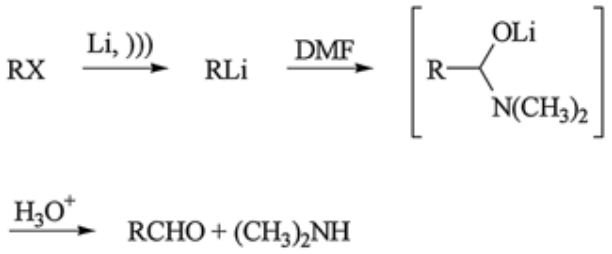

(18)

This special case of a Barbier reaction shows no reaction at an ultrasonic frequency of 50 kHz in diethyl ether, but reasonable yields at 500 kHz. This unexpected frequency effect can not be explained by cavitation and remains one of the open questions of sonochemistry.

Magnesium is used widely in the synthesis of Grignard reagents. Fast reactions and virtual absence of induction times under sonication allow safe and easy production of these valuable intermediates [21], [22].

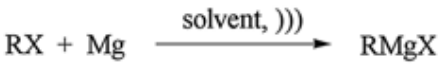

(19)

Sonication experiments on 2-bromobutane revealed a 60-fold increase in reaction rate even in impure wet ether [23]. Bogdanovic et al. developed a method to prepare a highly reactive magnesium-anthracene complex. Using this activated magnesium Mg*, a variety of Grignard reagents can be produced:

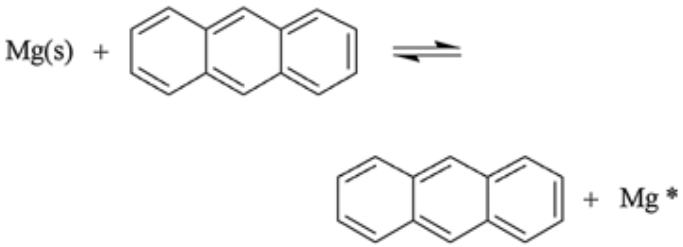

(20)

Grignard reagents are often used for transmetalation reactions to prepare less readily accessible organometallic compounds such as organoboranes and organozinc compounds.

Aluminum. Improvements in reaction rate and yield have been reported for the preparation of alkyl aluminum halides. An interesting result was the reaction of bromoethane with Al in THF to give ethylaluminum sesquibromide under ultrasound, which is not observed under stirring conditions.

Zinc. The direct formation of organozinc compounds from zinc and halides is possible, but transmetalation with Li is more reliable:

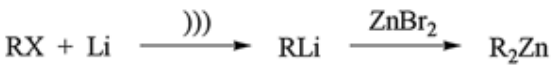

(21)

The Simmons-Smith cyclopropanation of alkenes with diiodomethane or dibromomethane and Zn benefits from ultrasound. Whereas under ultrasound 98 % yield after 2 h of irradiation are measured, only 50 % yield is possible in the stirred reaction.

A sonochemical variant of the Reformatzky reaction for the synthesis of β-hydroxyesters or β-lactams has been reported to give excellent yields compared to traditional methods using activated zinc or cosolvents. Allylations of carbonyl compounds

can be performed in an H_2O/THF/Zn slurry. Substantial rate enhancements are observed in the dehalogenation of aryl halides with zinc in aqueous HMPA with nickel(II) chloride. Cross coupling of perfluoroalkylzinc reagents with halides is possible.

Copper. Copper can be activated in an acoustic field to allow reaction at moderate temperatures. The Ullmann coupling reaction requires high temperatures. Sonicated copper is fragmented and cleaned to give highly reactive fresh surfaces and a 64-fold increase in reaction rate:

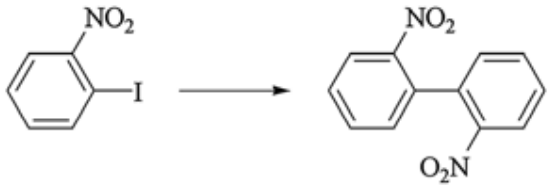

(22)

Nonmetals. In most reactions of nonmetals suspended in a liquid phase, the solid phase serves as one of the reagents and is consumed. Low solubility and slow reaction rates are common. Under ultrasound, a quasihomogeneous mixture is created by intense particle fragmentation and high mass transfer coefficients. Strong bases like KOH or NaOH are readily fragmented and undergo very fast and selective reactions. Oxidations of alcohols with potassium permanganate in benzene do not need water as a cosolvent under sonication. ANDO et al. [7] were able to activate manganese dioxide in a presonication process to give a material which oxidizes several alcohols. Quite unusual reductions of heteroatom-halogen bonds have been reported. Lithium aluminum hydrides are used to substitute aromatic halides in excellent yields.

CONTAMINE et al. [34] examined the Michael addition of a diethyl malonate anion to the double bond of phenyl styryl ketone. Solid KOH is used to promote the reaction in toluene. Two rate-limiting steps, the formation of a carbanion and its addition to the double bond, which is interpreted as a single electron transfer (SET), are likely. Under ultrasound, very fast reaction is observed and explained by enhanced mass transfer and KOH fragmentation, removing the first limiting step, and the influence of ultrasound on SET reactions.

3.1.2. *Catalytic Reactions*

Catalyst Preparation. Sonication has been applied to the preparation of catalysts for a wide range of processes. Hydrogenation catalysts like Pd/C, Pt/C, and Rh/C show better reactivity after mechanical activation and higher specific surface areas due to cleavage of the support. These activated catalysts are used in hydrogenation of alkenes, acetophenone, methylstyrene, and fat hardening. A remarkable increase in activity was found by SUSLICK [25], who sonicated Ni powder for up to 1 h in 1-octene. The metal showed a very clean surface (removal of Ni oxide) and a 10^5-fold activity, like freshly prepared Raney Ni, but a smaller specific surface area after prolonged sonication due to agglomeration.

Unusual catalysts are prepared by sonication of solutions of metal carbonyls in alkanes. Fe, Ni, Co, and Mo carbonyls give nanostructured amorphous metal or metal carbide powders.

Sonicated Catalytic Reactions. Inorganic supports like alumina, silica gel, and zeolites promote nucleophilic substitution with alkali metal cyanides in organic solvents. ANDO et al. [26] discovered the sonochemical switching of a reaction pathway in the reaction of benzyl bromide with potassium cyanide and alumina in toluene. While stirring leads to a mixture of the Friedel-Crafts products, benzyltoluenes, ultrasonic irradiation promoted nucleophilic substitution to give benzyl cyanide. It was concluded, that cyanide had poisoned the active Lewis acid sites, therefore decreasing the catalytic activity of the alumina.

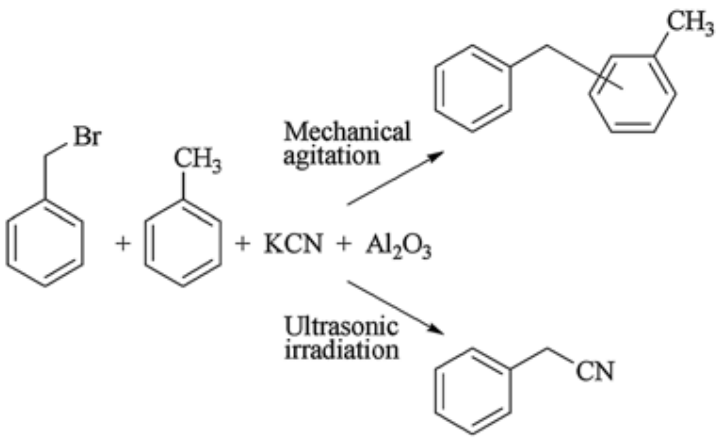

(23)

Hydrogenations of alkenes by formic acid or hydrazine and palladium-carbon with in situ generation of hydrogen have been reported [9]. Commercial applications such as the hydrogenation of soybean oil and fat hardening in three-phase processes have already found industrial utilization.

3.1.3. *Precipitation and Crystallization*

The application of ultrasound influences the nucleation, agglomeration, and growth of particles within the liquid phase. The main aim of the application of ultrasound in crystallization is the formation of a narrow particle size distribution [3], [35]. The sonication induces the homogenous nucleation in the sonicated volume [36], [37]. The nuclei are supposed to be of the same size, which is an advantage to the conventional seeded crystallization, which uses a seed suspension. This seed suspension

itself possesses a nonuniform particle size distribution, which is transferred to the crystallized product. High-power ultrasound promotes and improves nucleation and growth of crystals by cavitation bubbles, the disruption of seeds, and the breakage of larger crystals. There is some evidence that the sonication of supersaturated solutions leads to creation of highly uniform nucleation sites by disturbing the metastable equilibrium. The enhanced micromixing avoids nonuniform concentration gradients in the solution. Further on, sonocrystallization is applied to control the shape and structure of a crystal, called polymorphic controll. This plays an important role in the field of soft, i. e., biological materials like proteins or fatty acids [38]. Industrial applications of homogenous nucleation due to ultrasound are found in the frozen food industry, where even seeding prevents the disruption of cellular materials on freezing.

Precipitation is primarily controlled by the mixing of two solutions. The more homogenous the mixture is the smaller particles can be synthesized [39]. The application of ultrasound provides the possibility of intense and controlled mixing in the microscale. In contrast to flow reactors like T-mixer, rotor–stator mixer, and impinging jet mixer, the mixing with ultrasound is decoupled from the flow field. The specific power induced for mixing can directly be controlled by the power consumption of the ultrasound generator. Cavitation allows local high-power dissipation and with it intense micromixing. The process usually is designed that the reactants are directly injected into the cavitation zone. A second method uses a reactant like urea, which decomposes during sonication, to induce the precipitation process.

Particle synthesis due to the reduction of dissolved ions also applies ultrasound. The synthesis of gold nanoparticles [35] uses the presence of radicals, which in a second step act as a reducing agent to form the metal nanoparticles. The formation rate of the metal nanoparticles can be influenced by additives, which act on the radical formation rate.

3.1.4. *Atomization*

The production of atomized sprays from a liquid is facilitated by the use of ultrasonic devices. Normally, sprays are produced by forcing high-velocity liquids through a small aperture. Blockage and nonuniform size distributions are often observed. The initial velocity of the produced spray tends to form clusters. Ultrasonic devices make use of high velocity changes on a vibrating tip. The particle size can be controlled by either the frequency or the intensity of the ultrasonic horn. Materials like molten glass or metal are easily atomized, show a very narrow size distribution, and have a very low particle velocity.

3.1.5. *Disaggregation – Comminution*

The governing effect on size reduction is cavitation. Agglomerates of lower stability also are affected by the pressure waves of the acoustic field. The process parameters and the theoretical description derive from the conventional milling theory [40], [41].

$$x_{50,3} = C \cdot (E_v)^{-b} \tag{24}$$

where $x_{50,3}$ is the mean volume particle size in nm, C is a characteristic constant, E_v is the specific energy (kJ m^{-3}). The exponent b has to be drawn fron the plot $x_{50.3}$ versus log(E_v).

The process depends on the specific energy consumption, the specific power input of the transducer and the residence time.

Ultrasonic disagglomeration can be combined with conventional milling equipment. Sonication after a communution process is able to downsize agglomerates formed directly after the mechanical milling or agglomerates damaged by the mechanical stress [42].

3.1.6. *Cell Disruption*

The main effects of ultrasound which occur during biological applications are cavitation and heating of the medium. The heating is an immanent effect due to the absorption of the wave energy within the medium [43]. Considering biological molecules (proteins), their high sensitivity to thermal changes has to be taken into account (see ⟶ Enzymes – General Production Methods, Table 2). The application of ultrasound in biological systems usally is accompanied by an intense cooling. The heating depends of the distribution of the ultrasonic waves into the medium and its heat-transfer properties. Heat also is generated during cavitation. Because of the inhomogenity of biosystems, sufficient nuclei for bubble formation are present.

The mechanical action of cavitation typically causes cell lysis, disintegration, or defragmentation. At lower ultrasonic intensity, just above the cavitation threshold [44], subtle damages of the cell membrane occur, which lead to an increased permeability of macromolocules through the membrane. This effect is called sonoporation or permeabilization. Sonoporation is used to load cells with large molecules, i.e., DNA [45] or drugs.

With further increasing ultrasound intensity, the cavitation intensity and with it the mechanical stress acting on the cell membrane increases as well. The duration of sonication of a cell suspension can influence the quality of the obtained biosuspension. With increasing sonication, the activity of liberated proteins can decrease [46]. The main process parameters for cell disruption are: ultrasound intensity, duration of sonication, process temperature, protein/enzyme yield, protein/enzyme activity.

3.2. Ultrasonic Extraction

Ultrasonic extraction processes from biological cells, bacteria, microorganisms, or plants are used in research as well as on an industrial scale [5]. Cell disruption, enhanced micromixing, and impregnation of material by liquid microjets reduces processing time and allows less harmful solvents to be used for extractions. The normally lower temperatures allow very mild processing with high yields of alkaloids, essential oils, glycosides, and fragrances.

3.3. Liquid – Liquid Reactions

Only a few examples of the influence of ultrasound on liquid–liquid reactions have been reported in the literature. Most of

them benefit from the emulsifying effect of ultrasound and the resulting high interfacial area. In some cases, a phase-transfer catalyst became unnecessary. Most reactions are between immiscible aqueous and organic liquid phases.

The decomposition of organic halides is faster due to the better mass transfer from the organic phase into the aqueous phase, where secondary radical reactions with hydroxyl radicals lead to unspecific degradation. Examples of ester hydrolysis, substitution reactions, and addition reactions in two phase liquid–liquid systems can be found in [7].

Ultrasonic Mixing. Nonreactive mixtures of liquids can be mixed by ultrasound. High shear velocities and oscillating and transient bubbles create very intense local conditions, which are employed for homogenization or emulsification. Many industrial processes in the food and drug industries use the liquid whistle, a long-known device which uses a high-velocity liquid stream impacting a steel blade to create vibrations. Typical applications are the homogenization of water-oil mixtures for the production of soups, sauces or the rapid mixing of fruit juices from concentrates [3].

Emulsification. The main mechanisms for emulsification applying ultrasound are pressure gradients, cavitation and the resonance of capillary waves. The pressure gradient, Δp_{dr}, in a cavitation-free fluid has to overcome the capillary pressure, Δp_{cap}, of the droplet with the radius R_{dr} and the interfacial tension σ_L.

$$\Delta p_{dr} = \left.\frac{dp}{dy}\right|_{max} \cdot 2 \cdot R_{dr}$$
$$\Delta p_{cap} = \frac{2 \cdot \sigma_L}{R_{dr}} \tag{25}$$

Further, the frequency f_{us} of the ultrasonic field determines the deformation time t_{def}, which is available to perform the droplet breakage.

$$t_{def} = \frac{\pi}{f_{us}} \tag{26}$$

The critical deformation time [47-49] amounts to

$$t_{def} = \frac{\eta_{L,2}}{\sigma - \Delta p_{cap}} \tag{27}$$

where $\eta_{L,2}$ is the dynamic viscosity of the second, dispersed liquid. In the case of drop breakage, the tension σ is equal to the external pressure gradient Δp_{dr}. The effect of the pressure gardient is quite small below 100 kHz. With increasing frequency, the pressure gradient Δp_{dr} increases and with it the possible deformation of the droplet.

Droplets formed from the collapse of waves have a diameter about the same size as the wavelength itself. Typical process parameters in emulsification lead to a wavelength of about 50 micron, which is about the size of droplets within a raw emulsion before sonication. Therefore the mechanism of collapse of waves is not the governing effect forming fine-disperse emulsions.

It is supposed that primarily cavitation provokes the dispersion of large droplets into smaller droplets forming a stable emulsion. The raw emulsion is a heterogeneous system, therefore the asymmetrical bubble oscillations and collapses in the vicinity of droplets lead to liquid microjets and shock waves. The microjets induce strong local friction forces and pressure gradients, which are supposed to be the governing effect in drop dispersion. These effects and their quantification are subject to actual research, because a multiplicity of the experimental results names cavitation as main effect in emulsification.

The process result for ultrasound emulsification can be plotted in a diagram of droplet diameter versus energy per volume. The curves for ultrasonic emulsification show the same behaviour as those for conventional emulsification equipment. The efficiency of ultrasound emulsification is higher in o/w-emulsions than in w/o-emulsions. This is due to the higher intensity of cavitation in the aqueous phase.

3.4. Phase-Transfer Catalysis

Alkylations are important reactions in heterocyclic chemistry and provide a simple means for introducing substituents. Phase-transfer catalysis (PTC) allows very fast and facile reactions when combined with sonication. The alkylation of indole and carbazole with alkyl and benzyl halides under liquid/solid PTC conditions in toluene with solid KOH gives higher yields and shorter reaction times than the stirred reaction [20]. Reaction times are shortened from several hours to less than one hour under ultrasound.

The preparation of azides from primary alkyl halides and aqueous sodium azide under PTC conditions is easier and cheaper under ultrasound and smaller amounts of PTC reagents can be used.

3.5. Sonoelectrochemistry

Deposition of Metals. Acoustic microstreaming and transient events can aid the deposition of metals in electroplating processes [27]. The very high mixing efficiency in cavitating sound fields is used to destroy the polarization barrier surrounding the electrode. Constant mechanical treatment by transient bubbles leads to surface hardening, less porosity, better brightness and to very good adhesive deposits for a variety of metals, including copper and gold.

Electro-Synthesis. Electrochemical methods are of widely recognized importance in the generation of intermediates which are difficult to prepare by other means [24], [28]. Problems arise due to mass transfer limitations, fouling of the electrode, or the evolution of gases. Applications in the Kolbe reaction and dechlorination of polychlorinated biphenyls can be found. Reductions of benzaldehyde and benzoquinone are reported in [29], in which a novel electrochemical reactor employs the

ultrasound source as one electrode for the reaction. The production of microemulsions and the dispersing effects of ultrasound are used where immiscible liquids have to be treated. Applications are found in wastewater treatment and other environmentally important processes. An effective removal of gases produced at an electrode can help to shift reversible reactions. Chlorine and hydrogen removal leads to very effective electrochemical cells. Transient bubbles and intense micromixing are responsible for the disintegration of diffusion layers, thus eliminating mass transfer limitations in electrochemical reactions. Cyclovoltammogrammic measurements of reversible redox reactions show very characteristic peak shapes in the silent case due to diffusion limitations. Under sonication, step responses are observed, and the underlying electrode kinetics can be measured without mass transfer limitations. Limiting currents which are ten times higher than in the silent reaction are common. Due to the profound influence of ultrasound on electrode kinetics, a change in the product spectrum of electrochemical reactions is sometimes observed.

[Top of Page]

4. Homogeneous Systems

4.1. Aqueous Liquids

The sonochemistry of water, often called sonolysis, is mainly determined by the production of radicals. Among the earliest observations in sonicated aqueous solutions was the liberation of molecular iodine from solutions of KI. The so-called Weissler reaction is one of the standard reactions for the chemical dosimetry of cavitating liquids. The primary step is the homolytic cleavage of water in the bubble. Saturating the water with tetrachloromethane facilitates the production of radicals:

$$H_2O \xrightarrow{)))} OH\bullet + H\bullet \qquad (28)$$

Hydrogen radicals react with water molecules to give further $OH^{\bullet}$ radicals and molecular hydrogen. The recombination of $OH^{\bullet}$ radicals at the bubble-liquid interface forms hydrogen peroxide. VON SONNTAG [30] estimated the $OH^{\bullet}$ concentration in the liquid boundary layer to be 1.2×10^{-2} mol/L. Hydroxyl radicals and chlorine radicals react with iodine to yield after recombination molecular iodine as a complex with I^-:

$$OH^{\bullet}/Cl^{\bullet} + I^- \rightarrow OH^-/Cl^- + I^{\bullet}$$

$$2I^{\bullet} + I^- \rightarrow I_3^- \qquad (29)$$

The existence of hydrated electrons under noble gases has been proposed by Margulis [6].

$$H^{\bullet} + OH^- \rightarrow H_2O + e_{aq}^- \qquad (30)$$

Volatile components in the water enter the cavitating bubble and are attacked by the $OH^{\bullet}$ radicals. Dissolved gases like nitrogen yield HNO_2, HNO_3, and smaller amounts of NH_3 and N_2O. Mixtures of N_2 and H_2 lead to the production of NH_3; mixtures of H_2 and CO lead to formaldehyde. A mixture of N_2 with CO, CH_4, or HCHO gives small amounts of various amino acids.

Organic components are attacked by $OH^{\bullet}$ and $H^{\bullet}$ radicals within the bubble or in the bulk liquid to give a wide range of secondary radical products. The irradiation of aqueous solutions of maleic acid with traces of Br_2 leads to a rapid isomerization to fumaric acid. The reaction is initiated by the cleavage of the bromide bond, the addition of the bromide radical to the double bond of maleic acid. Isomerization and subsequent liberation of bromide radicals lead to fumaric acid, which precipitates. The degradation of organic solutes in water is used in the wastewater treatment for removal of traces of halogenated aromatic and aliphatic compounds.

Inorganic compounds also undergo secondary radical reactions. Redox reactions with various ionic components are reported in [7].

4.2. Nonaqueous Liquids

Most nonaqueous liquids have a much higher vapor pressure than water, and this strongly cushions the transient bubble collapse. Therefore, early measurements that showed only very few sonochemical effects in water-free liquids are easily understood. Such liquids exhibit sonochemical effects such as bond cleavage only at low ambient temperatures. The sonolysis of alkanes [31], for example, yields nearly the same products (H_2, CH_4, 1-alkenes) as the high-temperature pyrolysis reaction. On irradiation halocarbons undergo homolytic halogen-carbon bond cleavage and give molecular chlorine and secondary radical products.

Another example for the preparation of organometallic complexes and the different product spectrum compared to thermolytic or photolytic pathways was presented by SUSLICK who sonicated iron pentacarbonyl in alkanes:

(31)

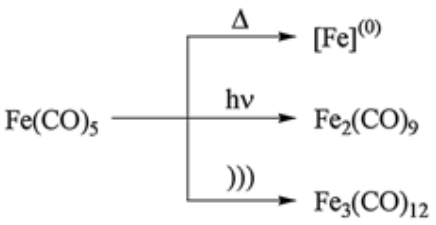

The products are decomposed to pyrophoric iron, when cavitation becomes too intense, and only solvents with higher vapor pressure and therefore less intense transient cavitation are suitable.

4.3. Polymers and Macromolecules

The preparation and degradation of long chain molecules under sonication is the result of two effects: creation of radicals by mainly mechanical chain cleavage of polymers and initiation of radical reactions by generation of primary radicals in transient bubbles.

Depolymerization. Early experiments with sonicated macromolecules revealed decreasing viscosity of the solutions. These changes can be attributed to a homolytic chain cleavage of the molecules. Drag forces acting on a dissolved polymer molecule are created in the vicinity of oscillating bubbles. High stresses are also created in shock waves from transient cavities. In dilute solutions, the polymer radicals react with solvent molecules to terminate the radical chain reaction. In concentrated solution, copolymerization and side chain branching occur. A lower limit for the molecular mass of the degraded polymer is observed in most cases. At a sufficiently small chain length, further degradation is stopped by the small drag forces acting on the molecule. The limiting molecular masses lie in the range of 10^4. The mechanical action of transient cavitation in polymer degradation becomes clear in the resulting molar mass distributions. Chains of higher mass are degraded more rapidly than those of lower mass. Degradation by ultrasound therefore leads to a narrower size distribution. Controlled depolymerization can be used to improve material properties such as mean polymer molecular mass, viscosity, and workability [3].

Polymerization. The creation of radicals in transient bubbles can be used to initiate radical polymerizations. Very high initial polymerization rates are observed at low temperatures for both pure monomers and monomers with initiators. Ultrasound produces a large number of radicals compared to traditional methods, and highly uniform size distributions are possible. With further sonication, depolymerization becomes increasingly important, and proper control of ultrasound parameters is necessary to achieve the desired polymer properties. The emulsification capabilities of ultrasound are utilized in suspension polymerization, where less surfactant is required and smaller size distributions are observed.

Copolymerization. The chain cleavage of polymers generates radicals which react with other reactants in the bulk. Concentrated polymer solutions (above 10 wt %) show side branching and copolymerization if two different polymers are present. Copolymerization of two polymers, for example, polystyrene and poly(methyl methacrylate), by ultrasound leads to random block and graft copolymers. In most cases the product is unfavorable, but the method can be a valuable tool for the preparation of mixtures from immiscible polymers. Copolymerization of a polymer in the presence of monomers leads to better defined block copolymers.

4.4. Sonoluminescence

Ultrasonic irradiation of liquids produces under certain instances light. This sonoluminescence is easily observed in water, ethylene glycol, and glycerin. In water, light emissions at 310 nm from excited $OH^{\bullet}$ radicals and a broad band in the entire continuum are observed. Whether this broad emission is due to black body radiation or excited molecules is still unknown. Other liquids show emission spectra from different excited species. Alkanes show a strong band in the C_2^* region (the Swan band). The production of C_2H_2 from sonicated alkanes provides evidence for excited species in the transient cavity. Sonochemically excited molecules and their reactions in the high-temperature region of bubbles seem to be responsible for the flame-chemistry-like behavior of sonoluminescence.

Single Bubble Sonoluminescence. In single-bubble sonoluminescence (SBSL) experiments [5], a bubble is trapped in a pressure antinode of a standing wave field. At a sufficiently high sound pressure of slightly above 0.1 MPa, this bubble oscillates stably but with high wall velocities and emits light in the UV/Vis region. Higher sound pressures cause a transient behavior and disappearance of the bubble. SBSL is used to examine the various parameters that influence bubble motion and sonochemical sonoluminescence efficiency.

Multi-Bubble Sonoluminescence. In extended sound fields, several types of bubbles, (e.g., stable and transient), are present. These create different temperatures and pressures inside oscillating bubbles and should therefore show a broader spectrum than in SBSL. MBSL experiments with aqueous solutions of various alcohols revealed a scavenging effect of these compounds. Sonoluminescence was suppressed at higher concentrations of alcohol, and this was interpreted in terms of reaction with $OH^{\bullet}$ and $H^{\bullet}$ radicals produced by homolytic water cleavage. Electrolytes can alter the SL spectrum. Emissions of neutral metal atoms indicate the possibility of metal vapor inside the cavities, which show the characteristic bands in the compression cycle when temperatures are high enough to vaporize the metals [5]. The mechanism of the generation of metal atoms from ionic metallic atoms by cavitation is still unknown, but theories about the reduction of the metal ion at the hot solid – liquid interface of a transient cavity seem plausible.

[Top of Page]

5. Equipment

5.1. Transducer Technology

Ultrasound transducers create the high frequency vibrations which are submitted into the liquid.

Two major types of transducers are used for liquids: liquid-driven and electromechanical.

Liquid-driven transducers are widely used in the food industry for mixing and homogenization. The principle is that of a liquid whistle. Liquid is pumped through a small orifice and hits a blade. The sudden pressure drop in the orifice causes hydrodynamic cavitation, and the blade generates ultrasonic vibrations and cavitation. A precise design of flow velocities, orifice and blade are necessary to allow high liquid throughputs with optimum mixing conditions.

Electromechanical transducers are based on piezoelectric or magnetostrictive principles. Both transducers need a high frequency generator for electrical supply. Magnetostrictive transducers are made of ferromagnetic material which alters its geometrical dimensions in an applied magnetic field. Bars or rods made of ferrite ceramics achieve high driving forces below 100 kHz. The major drawback of magnetostrictive transducers are their poor efficiency and rather broad frequency behavior.

Piezoelectric materials change their dimensions when charges are applied to opposite faces. These transducers have a natural resonance frequency at which the driving current produces the highest efficiency. Piezoelectric transducers are often made of barium titanate, lead metaniobate, and lead zirconate titanate ceramics and are supplied in most laboratory and industrial equipment. Operation at highest efficiency is responsible for the fixed frequency of common ultrasonic systems and the reason for the difficulty in examining the influence of frequency on sonochemical effects. Modern piezoelectric transducers have a sandwich structure with two electrically opposed piezo ceramics between two metal blocks. The sandwich is clamped together to prevent mechanical damage due to stress in the ceramic. The transducer usually has a length of half a wavelength, operates in compression mode, and generates amplitudes of up to 20 µm [32]. Piezoelectric transducers are depicted schematically in Figure 7.

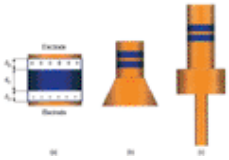

Figure 7. a) Schematic sketch of a piezoelectric ceramic with electrodes which changes its thickness d_e with applied charge; b) Transducer sandwich with two piezoceramics for low intensity devices; c) Transducer with booster and stepped horn

5.2. Reactors for Sonochemical Reactions

One of the major drawbacks in industrial applications of ultrasound is the complex scale-up procedure. Readily available laboratory equipment such as ultrasonic baths, high-intensity disintegrator horns, and cup horn reactors are used to measure sonochemical effects. All influencing parameters such as intensity, frequency, temperature, and vessel geometry must be the same for reproducible results. Critical ultrasonic parameters include the amplitude of the transducer, the ultrasonic intensity, the total power input, the specific power input per volume, the gas content, the local sound-energy distribution, and, in the case of heterogeneous reactions, the distribution of reactants. A successful scale-up must take these various factors into account.

5.2.1. *Laboratory Equipment*

Low-intensity applications can be tested in a commercial ultrasonic bath (Fig. 8 a). Several transducers are mounted at the bottom of a bath and are driven with the same frequency. A complicated standing wave field is created, and the sonochemical effects depend on the location of the samples in the bath and the filling height.

Probe systems are applicable in high intensity experiments at up to ca. 100 W/cm^2 at frequencies below 100 kHz. Multiplication of the transducer amplitude by a booster and a horn make it possible to increase the amplitude by a factor of around 10. The vibrating tip of the horn is dipped into the liquid and creates intense cavitation in its vicinity (Fig. 8 b). Ultrasound is attenuated rapidly, and only a few centimeters are sonochemically active. Such probes generate a large amount of streaming, so external stirring is unnecessary. The intense cavitation causes erosion of the tip and therefore contamination of pure liquids. A variant of the probe system is the so-called cup-horn reactor with indirect sonication of a submersed sample through a sound-transmitting liquid, which can be cooled (Fig. 8 c).

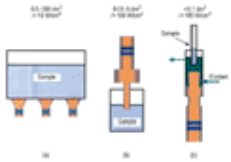

Figure 8. Laboratory equipment for sonochemical experiments a) Ultrasonic bath; b) direct-immersion horn; c) Cup-horn reactor

5.2.2. *Large-Scale Equipment*

Liquid-whistle reactors (Fig. 9 a) are still very important in homogenization and mixing processes. High throughput, stable operation, and the ability to adapt to existing flow systems are only some of the advantages of this system [23]. Fruit juices, sauces, and mayonnaise are prepared in large quantities. This low-intensity equipment can not be used where more severe cavitation is required. Other low-intensity devices are reactors with external or submersed plate or rod transducers (Fig. 9 b, c) which are easily fitted to existing tanks of several cubic meters and allow a fast scale-up. Tube transducers can also be retrofitted to existing plants. All low-intensity devices transmit sound energy over large volumes with dimensions greater than the liquid wavelength. The typical power input is roughly 50 W per transducer at a frequency of 20 kHz and a diameter of

6 cm.

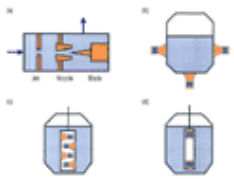

Figure 9. Low-intensity ultrasonic devices a) Liquid whistle; b) Tank reactor with wall-mounted transducers; c) Tank with submersed plate transducer; d) Tank with submersed rod transducer

High-intensity devices are applicable when the dimensions of the sonicated volume are equal to or smaller than the liquid wavelength. Probe systems can be used in small flow cells, which consist of several probes in series and can be retrofitted to existing plant. Gap reactors force liquids through the most active zone of a probe transducer. The power input per transducer reaches 2 kW with a radiating surface diameter of up to 10 cm.

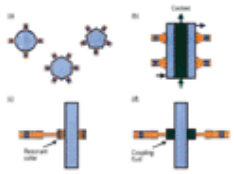

Figure 10. Tube reactors a) Conical, hexagonal, and heptagonal reactors; b) Raiganis tube reactor with axial cooling; c) Sodeva sonitube with stepped horn and resonant collar; d) Branson sonochemical reactor with coupling fluid

Another way to introduce high-intensity ultrasound are tube reactors (Fig. 10). Transducers are mounted on the outer wall of a tube and create intense cavitation in the center. Variants of this principle include the cylindrical pipe reactor; the cylindrical reactor with core cooling; hexagonal and heptagonal tubes, which allow easier transducer mounting on their plane walls; the Branson sonochemical reactor with a coupling fluid between the horn and the reactor wall; and the Sodeva sonitube with a stepped horn and a resonant collar mounted on a cylindrical pipe.

Gap reactors (Fig. 11 a) consist of a high-intensity stepped horn which is coupled with a small gap of a few millimeters. The fluid is driven through the gap and is treated by high-intensity ultrasound for a very short time. Particle size reduction and mixing are the main applications for these devices.

The reverbatory ultrasonic reactor (Fig. 11 b) consists of two opposing plates with a gap of a few centimeters. The use of two different frequencies (20 and 15 kHz) is very effective in particle size reduction. Other high-intensity systems are implemented in a tube (Fig. 11 c, d). Radial oscillations are responsible for the transmission of sound into the liquid. Examples for these devices are the Martin-Walter push-pull reactor and the Telsonic rod-reactor.

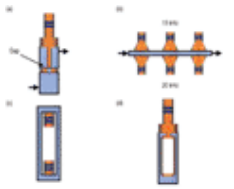

Figure 11. a) Gap reactor; b) Lewis reverbatory ultrasonic reactor; c) Martin Walter push-pull reactor; d) Telsonic rod reactor

5.3. Mapping of Sonochemical Reactors

The mapping of sononochemical reactors shows the intensity distribution of the pressure field within the reactor volume. This knowledge is essential, because most sonochemical effects need a minimum intensity of the ultrasound field. In the remaining reactor volume, where the sound intensity is too low, undesired side reactions can occur, decreasing the product quality.

5.3.1. *Hydrophone*

The hydrophone records the frequence of pressure changes within the sonicated liquid phase. With the help of a Fourier transformation (FFT), the frequence spectrum is calculated from the raw data. The response of the hydrophone clearly shows whether cavitation ocurrs or undisturbed pressure waves run through the liquid. Non-cavitating liquids only show the hydrophone response at the driving frequency of sonication. Cavitaton additionally generates overtones, thus the FFT shows peaks at multiples of the driving frequency [50].

The hydrophone has to be protected from cavitation effects. These lead to decalibration or even worse to destruction of the sensor. Because of its quite small size, the hydrophone can be moved within the sonicated vessel to measure the local intensity and frequency of the acoustic field.

5.3.2. *Thermocouples*

The measuring principle uses the heat generation during the absorption of ultrasound. The sensor consists of two thermocouples [48], [51]; one is covered with an absorbing material like silicon rubber. The temperature difference in the steady state of operation can directly quantify the intensity of the ultrasound field. Because of the quite small construction of thermocouples, an excellent local resolution of the sound field is possible. The precision of the measurement is negatively influenced by cavitation, which changes the frequency of the acoustic field.

5.3.3. *Optical Methods – Bubble Dynamics*

By applying of optical methods, it is possible to localize the cavitation zone; the optical properties of the liquid change when gas bubbles are created [48]. With the help of a high-speed video system, it is possible to track gas bubbles within or near the cavitation zone. This method can be used to localize secondary vortices in the flow field outside the cavitation center.

More precise is the particle image velocimetry (PIV) [52]. This method takes several photographs of the trajectories of small tracer particles in suspension in a thin plane of fluid, which is illuminated with a spread laser beam. The two components of the velocity vector in the illuminated plane are given by the traces left on the pictures during the exposure time. In cavitating liquids, the cavitation bubbles themselves can serve as tracers.

5.3.4. *The Weissler Reaction [52]*

Bubble cavitation leads in aqueous liquids to the formation of reactive radicals such as OH·, H·, and H_2O_2. These short-lived species are capable of inducing secondary oxidation and reduction reactions. For example, iodide (KI) can be sonochemically oxidized to triiodide by OH· radicals or H_2O_2. From aqueous solutions containing chlorocarbons, Cl· and Cl_2 are also liberated in high yields, which increases the rates of iodide oxidation. The rate of triiodide formation is easily monitored spectrophotometrically. This so-called Weissler reaction is still one standard dosimeter for sonochemical reactions because of its exclusive direct reaction to cavitation and its independence to temperature or sheer effects.

Because of mixing effects in the liquid, the Weissler reaction gives an integral value of the cavitation intensity. If the local intensity is of interest, the KI solution has to be separated from the main liquid in a special container, which is acoustically coupled to the liquid in the reactor [50].

[Top of Page]

References

General References

1. T. G. Leighton: *The Acoustic Bubble*, Academic Press, New York 1997.
2. F. R. Young: *Cavitation*, McGraw-Hill, Maidenhead 1989.
3. T. J. Mason, J. P. Lorimer: *Sonochemistry: Theory, Applications and Uses of Ultrasound in Chemistry*, Ellis Horwood Limited, New York 1989.
4. T. J. Mason: *Chemistry with Ultrasound*, Elsevier Applied Science, New York 1986.
5. K. S. Suslick: *Ultrasound: Its Chemical, Physical, and Biological Effects*, VCH Verlagsgesellschaft, Weinheim, Germany 1988.
6. M. A. Margulis: *Sonochemistry and Cavitation*, Gordon and Breach, Amsterdam 1995.
7. C. Einhorn, J. Einhorn, J. L. Luche, “Sonochemistry - The Use of Ultrasonic Waves in Synthetic Organic Chemistry”, *Synthesis* **11** (1989) 787 – 813.
8. P. W. Cains, P. D. Martin, C. J. Price, “The Use of Ultrasound in Industrial Chemical Synthesis and Crystallization. 1. Applications to Synthetic Chemistry”, *Org. Process Res. Dev.* **2** (1998) 34 – 48. Links
9. J. Lindley, T. J. Mason, “Sonochemistry Part 2 — Synthetic Applications”, *Chem. Soc. Rev.* **16** (1987) 275 – 311. Links

Specific References

10. R. E. Apfel, “Acoustic Cavitation Prediction”, *J. Acoust. Soc. Am.* **69** (1981) 1624 – 1633.
11. P. K. Chendke, H. S. Fogler, “Second-Order Sonochemical Phenomena — Extensions of Previous Work and Application in Industrial Processing”, *Chem. Eng. J.* **8** (1974) 165 – 178. Links
12. M. S. Plesset, R. B. Chapman, “Collapse of an Initially Spherical Vapour Cavity in the Neighbourhood of a Solid Boundary”, *J. Fluid Mech.* **47** (1971) 283 – 290.
13. I. Hansson, V. Kedrinskii, K. A. Moerch, “On the Dynamics of Cavity Clusters”, *J. Phys.* **D15** (1982) 1725 – 1734.
14. E. A. Neppiras, “Acoustic Cavitation”, *Phys. Rep.* **61** (1980) 159 – 251.
15. S. W. Woodside, B. D. Bowen, J. M. Piret, “Measurement of Ultrasonic Forces for Particle-Liquid Separations”, *AIChE J.* **43** (1997) 1727 – 1736.
16. K. A. Kusters, S. E. Pratsinis, “Ultrasonic Fragmentation of Agglomerate Powders”, *Chem. Engng. Sci.* **48** (1993) 4119 – 4127.
17. Y.-S. Chen, “Zur Frage des Einsatzes von Hoch-Intensitäts-Ultraschall in mehrphasigen Systemen”, Dissertation, TU Clausthal, 1997.
18. F. Contamine, F. Faid, A. M. Wilhelm, J. Berlan, H. Delmas, “Chemical Reactions under Ultrasound: Discrimination of Chemical and Physical Effects”, *Chem. Engng. Sci.* **49** (1994) 5865 – 5873.
19. S. J. Doktycz, K. S. Suslick, “Interparticle Collisions Driven by Ultrasound”, *Science* **247** (1990) 1067 – 1069. Links
20. Y. Goldberg, R. Sturkovic, E. Lukevics, “Heterolytic Sonochemistry”, *Heterocycles* **29** (1989) 597 – 627. Links
21. C. Horst, U. Kunz, U. Hoffmann, “Design, Modeling and Performance of a Novel Sonochemical Reactor for Heterogeneous Reactions”, *Chem. Engng. Sci.* **51** (1996) 1837 – 1846.
22. C. Horst: “Ultraschallreaktoren zur Durchführung von heterogenen Flüssig-Feststoffreaktionen: Entwurf, Betrieb und Modellierung einer Versuchsanlage zur Darstellung von Grignardverbindungen”, Dissertation, TU Clausthal, 1997.
23. A. B. Pandit, V. S. Moholkar, “Harness Cavitation to Improve Processing”, *Chem. Engng. Prog.* **7** (1996) 57 – 96.
24. R. G. Compton, J. C. Eklund, S. D. Page, “Sonovoltammetry: Heterogeneous Electron-transfer Processes with Coupled Ultrasonically induced Chemical Reaction. The “Sono-EC” Reaction”, *J. Phys. Chem.* **99** (1995) 4211 – 4214.
25. K. S. Suslick, D. J. Casadonte, “Heterogeneous Sonocatalysis with Nickel Powder”, *J. Am. Chem. Soc.* **109** (1987) 3459 – 3461. Links

26. T. Ando *et al.*, “Sonochemical Switching of Reaction Pathways in Solid-Liquid Two-phase Reactions”, *J. Chem. Soc. Chem. Commun.* (1984) 439 – 440.
27. M. J. Ashley, “Ultrasonics in Chemical Processing”, *The Chemical Engineer* (1974) 368 – 371.
28. D. J. Walton, L. D. Burke, M. M. Murphy, “Sonoelectrochemistry: Chlorine, Hydrogen and Oxygen Evolution at Platinised Platinum”, *Electrochim. Acta* **41** (1996) 2747 – 2751. Links
29. A. Durant, H. Francois, J. Reisse, A. Kirsch-DeMesmaeker, “Sonoelectrochemistry: The Effects of Ultrasound on Organic Electrochemical Reduction”, *Electrochim. Acta* **41** (1996) 277 – 284. Links
30. C. von Sonntag: “Aqueous sonochemistry: free radicals in two phases”, Congress Proceedings of the 6th Meeting of the European Society of Sonochemistry, May 10 – 14, 1998, Rostok – Warnemünde, Germany.
31. K. S. Suslick, P. F. Schubert, “Sonochemistry of $Mn_2(CO)_{10}$ and $Re_2(CO)_{10}$”, *J. Am. Chem. Soc.* **105** (1983) 6042 – 6044. Links
32. A. Stroh, “Technical Considerations in Applying High Intensity Sound to Physicochemical Processes”, *Chem. Engng. Prog.* **109** (1971) 128 – 134.
33. L. A. Crum, “Bjerknes Forces on Bubbles in Acoustic Fields”, *J. Acoust. Soc. Am.* **57** (1975) 1363 – 1370.
34. F. Contamine, A. M. Wilhelm, J. Berlan, H. Delmas, “Activation of a Solid-Liquid Chemical Reaction by Ultrasound”, *Chem. Engng. Sci* **50** (1995) 554 – 558.
35. Y. Mori, N. Kitamoto, K. Tsuchiya: “Effect of kind of alcohols as additives on preparation of gold nanoparticles by sonochemistry”, *J. Chem. Eng. Japan* **38** (2005) 283–288.
36. K. Ohsaka, E. H. Trinh: “Dynamic nucleation of ice induced by a single stable cavitation bubble”, *Applied Physics Letters* **73** (1998) 129–131. Links
37. Z. Guo, *et al.*,: “Effect of ultrasound on anti-solvent crystallization process”, *Journal of Crystal Growth* **273** (2005) 555–563. Links
38. S. Ueno, R. I. Ristic, K. Higaki, K. Sato: “In situ studies of ultrasound-stimulated fat crystallization using synchrotron radiation”, *Journal of Physical Chemistry B* **107** (2003) 4927–4935.
39. Y. Seida, T. Nakano, Y. Nakamura: “Crystallization of layered double hydroxides by ultrasound and the effect of crystal quality on their surface properties”, *Clays and Clay Minerals* **50** (2002) 525–532. Links
40. M. Pohl, S. Hogekamp, N. Q. Hoffmann, H. P. Schuchmann: “Dispersion and deagglomeration of nanoparticles with ultrasound”, *Chem. Ing. Tech.* **76** (2004) 392–396.
41. U. Teipel, K. Leisinger, I. Mikonsaari: “Comminution of crystalline material by ultrasonics”, *International Journal of Mineral Processing* **74** (2004) S183–S190. Links
42. Z. Herceg, *et al.*: “Fine milling and micronization of organic and inorganic materials under dynamic conditions”, *Powder Technol.* **139** (2004) 111–117. Links
43. D. Miller, S. Pislaru, J. Greenleaf: “Sonoporation: Mechanical DNA Delivery by Ultrasonic Cavitation”, *Somatic Cell and Molecular Genetics* **27** (2002) 115–134. Links
44. H. Y. Lin, J. L. Thomas: “Factors affecting responsivity of unilamellar liposomes to 20 kHz ultrasound”, *Langmuir* **20** (2004) 6100–6106. Links
45. M. Fechheimer, C. Denny, R. F. Murphy, D. L. Taylor: “Measurement of Cytoplasmic Ph in Dictyostelium Discoideum by Using a New Method for Introducing Macromolecules into Living Cells”, *European Journal of Cell Biology* **40** (1986) 242–247. Links
46. L. Benov, J. Al-Ibraheem: “Disrupting Escherichia coli: A comparison of methods”, *Journal of Biochemistry and Molecular Biology* **35** (2002) 428–431. Links
47. P. Walstra: “Principles of Emulsion Formation”, *Chem. Eng. Sci.* **48** (1993) 333–349. Links
48. O. Behrend: *Mechanisches Emulgieren mit Ultraschall*, GCA, Herdecke, 2002.
49. O. Behrend, H. Schubert: “Influence of hydrostatic pressure and gas content on continuous ultrasound emulsification”, *Ultrasonics Sonochemistry* **8** (2001) 271–276. Links
50. P. R. Gogate, P. A. Tatake, P. M. Kanthale, A. B. Pandit: “Mapping of sonochemical reactors: Review, analysis, and experimental verification”, *AIChE J.* **48** (2002) 1542–1560. Links
51. B. Pugin: “Qualitative Characterization of Ultrasound Reactors for Heterogeneous Sonochemistry”, *Ultrasonics* **25** (1987) 49–55. Links
52. A. Weissler, H. W. Cooper, S. Snyder: “Chemical Effect of Ultrasonic Waves – Oxidation of Potassium Iodide Solution by Carbon Tetrachloride”, *J. Am. Chem. Soc.* **72** (1950) 1769–1775. Links

[Top of Page]

see also ⟶ On-Line Monitoring of Chemical Reactions – Ultrasonic Methods

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Spray Forming of Metals

UDO FRITSCHING, Stiftung Institut für Werkstofftechnik, Bremen, Federal Republic of Germany
KLAUS BAUCKHAGE, Stiftung Institut für Werkstofftechnik, Bremen, Federal Republic of Germany

1. Background and Potential of the Process 1
2. Atomization 3
3. Spraying 9
4. Deposition 16
5. Integral Modeling 20
6. References 21

Symbols

A	m^2	nozzle exit area
C_i	-	empirical constants of compaction efficiency
d_l	m	liquid jet diameter
d_p	m	particle diameter
d_{32}	m	Sauter mean diameter
$d_{50,3}$	m	mass median diameter
f_L	-	liquid contents of particles
GMR	-	gas to metal mass flow ratio
I_p	m^2/s	particle momentum
K_{lub}	-	empirical factor, depending on atomizer geometry
k	-	deposition efficiency
k_1	-	empirical constant of mass flux distribution
Ma	-	Mach number
$\dot{m}_1$	kg/s	liquid mass flow rate
$\dot{m}_g$	kg/s	gas mass flow rate
$\dot{m}_{spray}$	$kgm^{-2}s^{-1}$	particle mass flux in spray
$\dot{m}_{compac}$	$kgm^{-2}s^{-1}$	deposited mass flux
$\dot{m}_{max}$	$kgm^{-2}s^{-1}$	maximum mass flux
Oh	-	Ohnesorge number
p_0	bar	stagnation pressure
p_1	bar	primary gas pressure
p_2	bar	atomization (secondary) gas pressure
R	$kJkg^{-1}K^{-1}$	gas constant
r	m	radius
$r_{0.5}$	m	half width of mass flux profile
T_0	K	stagnation temperature
T_L	K	liquidus temperature
T_p	K	particle temperature
t	s	time
u_{max}	m/s	maximum gas velocity
u_{rel}	m/s	relative velocity
We	-	Weber number
x	m	coordinate
μ_f	-	nozzle friction factor
λ	m	wavelength
ν_l	m^2/s	kinematic liquid viscosity
ν_g	m^2/s	kinematic gas viscosity
ρ_g	kg/m^3	gas density
ρ_l	kg/m^3	liquid density
σ	-	standard deviation
σ_l	N/m	liquid surface tension

Abstract

Spray forming is an advanced process route to produce near-net-shaped preforms from metallic materials. As it combines the main advantages of the classical processes of metal casting and powder metallurgy, its products are characterized by extraordinary material properties. Since the 1970s this process has been developed and it is now in its commercialization state in several applications, e.g., in the automotive and aerospace industry. Though essentially a metallurgical process, understanding and optimization is based on principles of process and chemical engineering. This is reflected by means of the three process stages as: melt atomization, spray evolution, and droplet deposition. In this contribution these three subprocesses will be introduced and basic features of the process will be discussed

1. Background and Potential of the Process

Spray forming is a metallurgical process which combines the main advantages of the classical processes of:

– Casting: near net shape manufacturing of large preforms, and
– Powder metallurgy: homogeneous grain structure for small components.

A schematic of a typical spray forming plant is shown in Figure 1. The metal melt is poured from the crucible into the tundish and flows into

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b24_b02.pub2

the atomization region where it is usually disintegrated by impinging high-energy inert gas jets. In the resulting spray cone, the molten metal droplets partly solidify and are accelerated towards the substrate, where they are compacted to a solid preform (deposit).

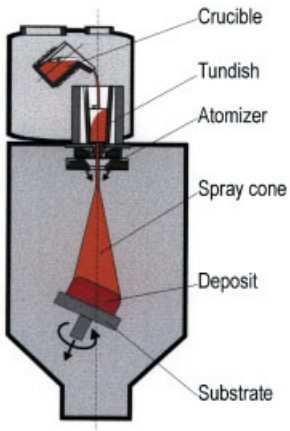

Figure 1. Schematic of the spray-forming process

The idea of spray forming was developed in the late 1960s by Singer [1] at the University of Swansea in Wales. In the 1970s spray forming was further developed for the production of thin metal sheets directly from the melt, as an alternative to the conventional metallurgical route of casting and subsequent forming of large slabs. Also in Wales, members of Singer's group commercialized the spray forming process by founding the company Osprey Metals Ltd. in Neath. Since then, research groups worldwide have taken an active part in the research and development of the fundamentals and potential applications of spray forming [2 – 8]. At present, numerous industrial applications of spray-formed materials are successfully at the market. Application areas are within automotive and aerospace, electronics and manufacturing. A review of the applications of spray forming is given in [9, 10].

The materials that are commonly used for production of semifinished products by spray forming are:

- Aluminum-, copper- and iron-base metals and alloys which are difficult to cast because they tend to segregate
- Superalloys (e.g., on nickel-base) and titanium
- Particle-reinforced metal-matrix composites (MMCs)

The main geometries which are produced by spray forming are shown in Figure 2.

Sheets are manufactured by scanning (oscillating) the atomizer or by using linear-slit nozzles to spread the spray cone over a certain area and by spraying onto a substrate in linear motion. Metal sheets are of great commercial interest, e.g., in the aluminum industry, but are difficult to spray form up to now due to processing problems.

Tubes and *rings* are produced by scanning the atomizer or by stationary spraying onto a tubular or cylindrical substrate. The core is drilled out or remains in the form to give a two-layered product (with different properties on the inside and the outside). Spray-formed tubes are used in the chemical industry, and spray-formed rings will have applications in turbines and propulsion units (aeronautics and astronautics).

Billets are presently the most common geometry manufactured by spray forming. Cylindrical billets are commonly used as preforms for a subsequent extrusion process. Billets are manufactured by off-center spraying at an angle of inclination onto a rotating cylindrical substrate, which is withdrawn during the process, generally in the vertical direction for a standing billet, but sometimes in the horizontal direction to produce a lying billet with larger dimensions. To increase the production rate and throughput, sometimes two atomization units are used simultaneously [11 – 13]. Spray-formed billets are presently produced up to a diameter of approximately 0.5 m and up to a height of 2 m. Applications are in automotive industries (aluminum) or in superconductors (copper).

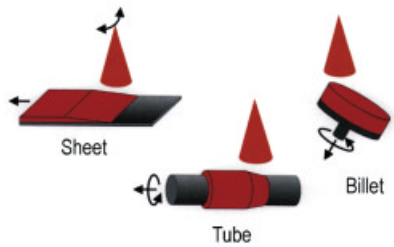

Figure 2. Spray forming of different preform geometries

The main advantages and properties of spray-formed products and materials are:

- Absence of macro-segregations
- Homogeneous, equiaxial grain structure
- Increased strength
- Increased elongation
- Low oxygen content
- Better hot workability

The specific advantages of different spray-formed materials, the properties of spray-formed near net shaped products, and the state of the art in the industrial application of the process are documented in the periodical international user conferences on spray forming (ICSF) [14 – 16], in the publication series of the scientific contributions on spray forming in the special research program at the University of Bremen [17 – 23], and in the Proceedings of the International Confereneces on Spray Deposition and Melt Atomization SDMA [24, 25], including contributions to fundamentals and applications in:

- Copper-base materials [26]
- Aluminum-base materials [27, 28]
- Ferrous alloys and steels [29 – 31]
- Titanium and ceramic-free superalloys [32]
- Compound materials and MMCs [33]
- Lightweight materials (e.g., magnesium) [34]

From a process engineering viewpoint, the sprayforming process can be divided into several sequential unit operations. Different setups and ways of describing and combining these fundamental process steps can be found in [4, 35 – 37].

Starting point of an analysis is the subdivision of the spray-forming process into three main stages [38], as shown in Figure 3.

1) Preparation of the melt and atomization: This stage includes the preparation of the alloy and the delivery and the disintegration of the fluid from the continuous melt flow to a spray with a droplet size distribution.
2) Particle transport in the spray cone: The behavior and development of the spray cone are described by the multiphase flow of metal droplets/particles and gas, and by momentum and heat transfer within and between phases.
3) Impact and deposition of droplets: The impact of the droplets, their compaction to give a deposit, the establishing of the so-called mixing layer (or mushy zone) at the deposit surface, and the behavior of the product are involved in this stage.

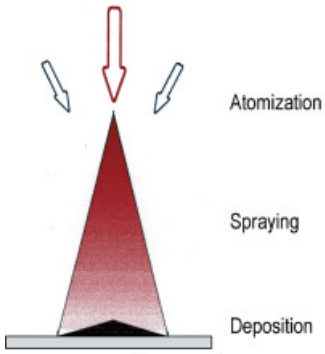

Figure 3. Three process stages of spray forming

The possibilities for an optimized, automated spray-forming process are discussed in [39]. Starting from the division into three main stages, one can relate the geometrical and operational parameters of the process to those stages which must be visualized, measured, and controlled. Changing the operational parameters of the process results in a complex behavior of the coupled stages of the spray-forming process and, ultimately, a quite different quality of the sprayed product. The control of the droplet size spectrum in the spray cone is the most important control parameter in a spray-forming process. The key parameter for online process control is regulation of the gas to metal flowrate ratio (GMR) by control of the gas mass flow rate of the atomizer (assuming constant thermal conditions). The aim is to influence the atomization process and therefore directly influence the coupled multiphase transport and solidification of particles in the spray cone. The disturbing variable of this control process is mainly the fluctuation of the metal melt flow rate in operation (due to changing melt properties during the process or stochastic influences), which must be continuously measured [40] and controlled during operation to give a closed control loop.

2. Atomization

Of the various atomization principles and devices which are used for disintegration of molten metals [41, 42] for powder generation and

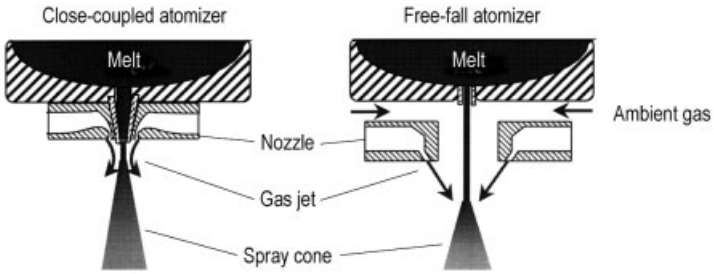

Figure 4. Close-coupled and free-fall atomizers for metal powder production and spray forming

spray-forming applications, twin-fluid atomization with inert gases is best suited [43], because:

- High mass flow rates of melt can be atomized to increase the yield
- High heat transfer rates from the droplets can be realized
- High momentum transfer rates can be achieved to accelerate the droplets
- Low oxidation rates of the material in the spray process are obtained

Common to all twin-fluid atomizers for molten metal disintegration are a vertical flow of a cylindrical melt jet from the tundish in the direction of gravity and a parallel or inclined, coaxial gas jet with high kinetic energy. Two kinds of atomizers can be distinguished (Fig. 4): the free-fall configuration and the close-coupled (confined) atomizer. The efficiency of the close-coupled atomizer is higher (in terms of smaller droplet sizes) compared to the free-fall atomizer, due to the closer spacing of the gas jets and the melt outlet. However, the close-coupled atomizer has the higher potential for melt freezing at the nozzle outlet, because the temperatures of the expanding pressurized atomizer gas are usually below 0 °C and high cooling rates of the melt flow result at the tip of the tundish. Freezing is a major problem, because the spray forming is a batch process, and freezing often occurs in the start-up phase.

Another advantage of the free-fall atomizer is the possibility of mechanically scanning or oscillating the separate atomizer gas ring. This gives the process operator an additional possibility for influencing the droplet mass flux distribution in the spray cone. This important parameter can normally be changed only over a narrow range for a given process by changing the atomizer gas pressure, which simultaneously changes the overall behavior of the spray (e.g., by changing the droplet sizes).

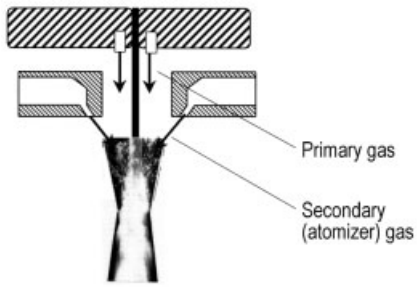

Figure 5. Two-stage atomization (primary and secondary gas flow) and particle recirculation in the free-fall atomizer

The analysis of the atomization process requires knowledge of the flow field of the atomizer gas in the vicinity of the twin-fluid nozzle [44 – 47]. The gas flow field is responsible for the melt-disintegration process, but also has a guidance and shielding function with respect to the initial behavior and distribution of the spray cone. These two requirements can be achieved by means of a two-stage free-fall atomizer (Fig. 5), in which a primary gas flow emerging close to the melt outlet guides the melt flow into the atomization region, while a secondary gas flow field emerging from the lower gas ring impinges on the central melt stream and disintegrates the melt [48]. Due to the inclination of the secondary (atomization) gas flow, recirculation areas are frequently observed in the vicinity of the nozzle. Individual particles entering these zones may be transported vertically upwards, in opposite direction to the main gas flow, as can be seen in Figure 5, in which a photograph of part of the atomization area is mounted. When these particles hit the nozzle body close to the melt outlet

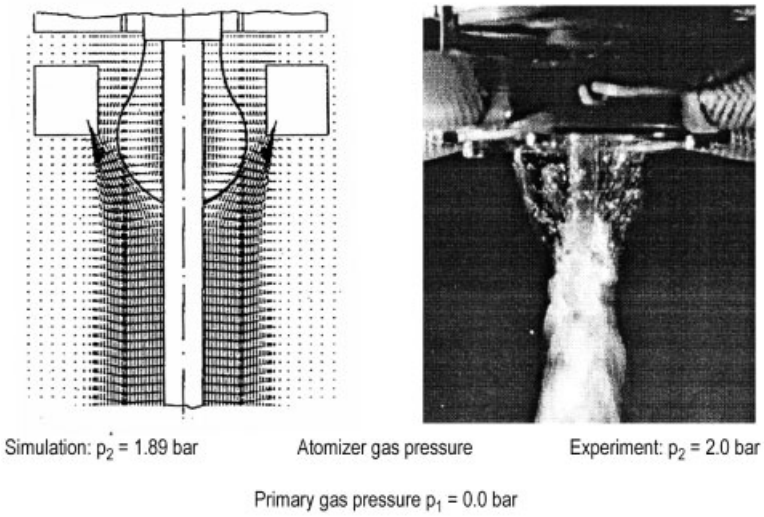

Figure 6. Comparison of simulation and experiment for twin-fluid atomization [49]

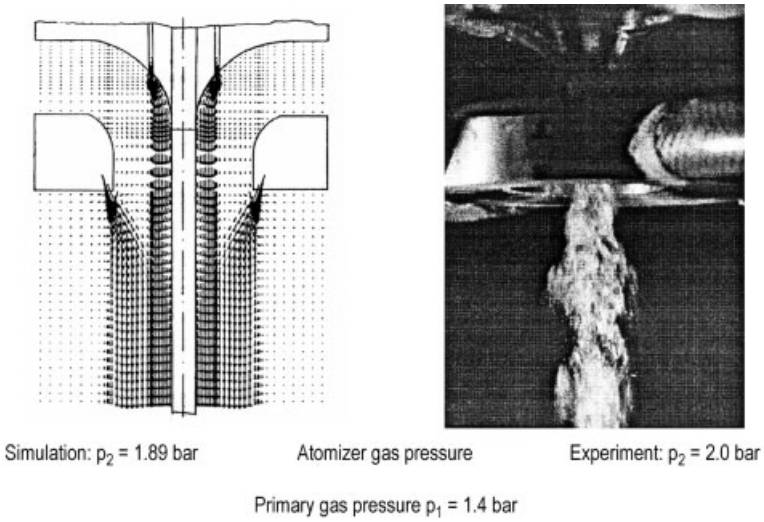

Figure 7. Comparison of simulation and experiment for twin-fluid atomization [49]

they may stick there, resulting in a severe disturbance of the process and in extreme cases, complete plugging of the nozzle.

Figure 6 compares a numerical simulation of the (subsonic, $Ma \leq 1$) gas flow in the vicinity of nozzle (vector diagram) with the atomization of a model fluid (water) in a free-fall atomizer [49]. The atomization gas pressure in this case is $p_2 = 1.89$ bar abs.; therefore, the gas exit condition is critical with an exit velocity equal to the velocity of sound ($Ma = 1$) but without any shock cells in front of the nozzle. In this analysis, no primary gas flow is applied ($p_1 = 0$). The secondary atomizer gas flow exits the gas ring with an inclination angle of $10\,^\circ$ and hits the centerline of the atomizer just above the theoretical atomization point (the intersection point of the atomizer centerline with the gas jet exit line). The deviation is due to the contraction of the atomizer gas flow. Ambient gas is entrained and accelerated into the inner flow field. This gas stream flows through the gap between the

main body of the nozzle and the lower atomizer gas ring into the inner area of the flow field. In the middle of the inner area, a large recirculating vortex flow can be seen, the border of which is indicated by a solid line in Figure 6. Here, upward vertical gas velocities of up to 50 m/s are attained. Individual particles entering this area of upward flowing gas are accelerated towards the main body of the nozzle and the melt exit. This is also observed in the experimental model atomization process: individual droplets which are transported vertically upwards can be seen above the atomization region.

An appropriate nozzle configuration and condition for a failsafe operation of the atomization nozzle in spray forming of molten metals needs:

- A smooth outer nozzle contour, which exactly follows the streamline structure to prevent flow separation at sharp nozzle edges
- Location of the primary gas outlet as close as possible to the melt flow
- A suitable gap width between the main body of the nozzle and atomizer gas ring
- A minimum amount of primary gas flow, in combination with the atomizer gas pressure

The calculated gas flow field in an optimized free-fall nozzle configuration is shown in Figure 7 for an atomization gas pressure of p_2 = 1.89 bar abs. and a primary gas pressure of p_1 = 1.4 bar abs. [49]. Here, the gas flow field in the vicinity of the nozzle has no recirculation zone. The experimental model atomization shows the successful prevention of droplet recirculations for this nozzle.

This underlines the importance of the primary gas flow. Also the geometric contour of the atomizer gas exit configuration must be analyzed. In a conventional free-fall atomizer, the gas flow exit configuration is a line of individual cylindrical holes without inner contour on a common circumference of the atomizer gas ring. The atomization gas pressure is typically between 3 and 8 bar abs. ($Ma > 1$). Therefore, the individual gas jets exit the atomization nozzle in a choked (underexpanded) condition with a local pressure above ambient pressure. This flow generally tends to form a series of shock waves and expansion cells that cause energy losses [50]. An investigation of the transonic atomizer gas flow behavior in the vicinity of the nozzle shows that the use of contoured gas nozzles (Laval type) to give supersonic gas outlet conditions ($Ma > 1$), significantly augments the kinetic energy in the atomization region [45]. This behavior is shown in Figure 8, for the simulated flow field of a free-fall atomizer operated at p_2 = 5 bar abs. The left half of the figure is the result for an underexpanded gas flow, and the right half for a fully expanded supersonic gas flow field.

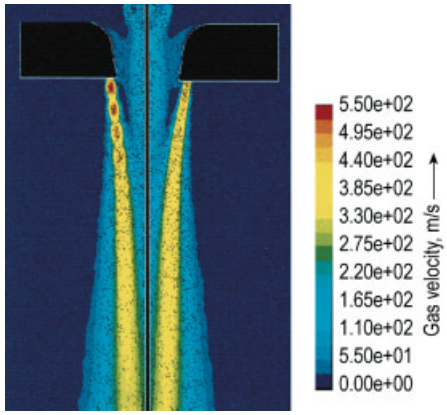

Figure 8. Numerical simulation of gas flow in the nozzle vicinity of a free-fall atomizer; underexpanded (left) and idealy expanded (right) [45]

The gas flow field disintegrates the melt by transfer of axial and tangential momentum to its surface, resulting in the formation of droplets and the spray cone. One of the main parameters describing the spray is the droplet size distribution. It is commonly described by an empirical correlation [52] for the mass median particle diameter as:

$$d_\mathrm{p} = d_\mathrm{l} K_\mathrm{lub} \left(\frac{\nu_\mathrm{l}}{\nu_\mathrm{g}} \frac{1}{We} \{1 + \frac{\dot{m}_\mathrm{l}}{\dot{m}_\mathrm{g}}\} \right)^{0.5} \tag{1}$$

where K_lub is an empirical factor which mainly depends on the atomizer geometry. The Weber number

$$We = \frac{\varrho_\mathrm{l} u_\mathrm{max}^2 d_\mathrm{l}}{\sigma_\mathrm{l}} \tag{2}$$

is calculated by means of the melt properties and the maximum gas velocity u_max of the atomizer gas flow field. The maximum gas velocity in the atomization region has been measured and correlated for the typical free-fall arrangement shown in Figure 7 as [45]:

$$u_\mathrm{max} = 645.25\, \dot{m}_\mathrm{g}^{0.585} \tag{3}$$

where the gas mass flow rate is related to the absolute atomization gas pressure in the choked flow regime by:

$$\dot{m}_g = 0.685\,\mu_f A\,\frac{p_0}{\sqrt{RT_0}} \tag{4}$$

where the friction coefficient μ_f is 0.84. The result of a sample calculation of the resulting mean droplet sizes for various atomized pure metals is shown in Figure 9.

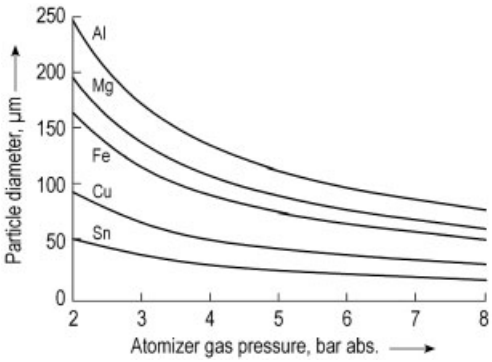

Figure 9. Particle diameter of melt droplets from twin-fluid atomization

A more detailed break-up model for description of melt-jet fragmentation needs the description of several subsequent stages of the atomization process as:

- the gas velocity field,
- the primary interaction of gas and melt jet and its break-up,
- the secondary interaction of gas and melt fragments and droplet break-up,
- drop collision effects as coalescence or fragmentation,
- the coupling of dispersed phase and continuous phase.

An uncoupled numerical computation of the atomization process and spray development is described here where the most relevant process parameters for atomization as the gas-to-melt mass flow ratio GMR are taken into account. In the computation, every droplet/particle in the flow field (no particle parcels) is tracked. The local acceleration of the droplet, the effect of turbulence, the secondary break-up, and the cooling of the molten droplet are taken into account. Drop collision effects are neglected.

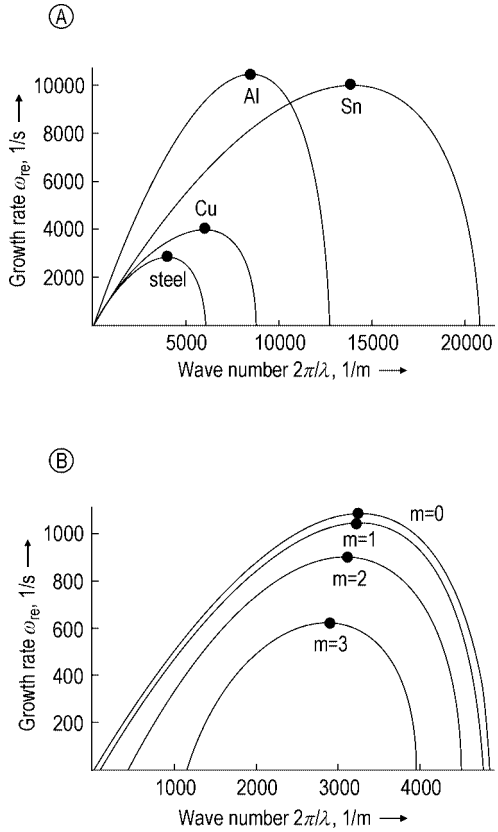

Figure 10. Stability curves for metal jet, $d_0 = 4$ mm, $u_0 = 2$ m/s, $u_{gas} = 100$ m/s (A), Sn, $u_{gas} = 50$ m/s (B) [55]

Theoretical investigations of the primary liquid jet break-up in a gaseous environment are based on the (linear) stability analysis of the coupled gas and liquid flow fields. For an inviscid flow, this is the classical Kelvin-Helmholtz instability as derived e.g. in [53, 54] for the description of the stimulation of the primary disintegration process in the area of small wavelengths within coaxial atomization of liquids.

Comparison of different modes of stability for a jet of molten metal in a coaxial gas flow has been discussed in [55]. Typical results for perturbation growth rates of different metal melts are shown in Figure 10 in the upper part and for the four basic modes of instability (n = 0, 1, 2 and 3) for a tin melt jet in Figure 10 in the lower part. The influence of material properties on melt-jet stability is mainly due to the surface tension, increasing from tin ($\sigma = 0.544$ N/m) via copper ($\sigma = 1.31$ N/m) to steel ($\sigma =$

1.83 N/m). Increasing values for surface tension result in decreasing Weber numbers, which result in smaller amplitudes of the growth rate. In the same way the dominant wavelength of jet disintegration increases. Higher-order instability modes are important. Maximum growth rates are observed for the symmetric mode m = 0, but the asymmetric growth mode m = 1 has almost identical growth rates while higher modes show somewhat smaller growth rates.

The secondary break-up process generally is defined as the fragmentation of ligaments into droplets, including the successive break-up of droplets into smaller droplets [56, 57]. In gas atomization of melts in free-fall atomizers, usually the melt mass flow rate (typically 0.05 ... 0.5 kg/s) is in the order of the gas mass flow rate. The describing parameter of the fragmentation process, the GMR, is about 1 to 2 and the atomization zone is characterized by:

- high acceleration of particles,
- high cooling rates and
- high particle bulk density.

At moderate superheats for metallic melts, the effect of viscosity is small. Different mechanisms of the break-up of molten metallic droplets were described by Krzeczkowski [58]. He distinguished between four break-up mechanisms:

- bag break-up for $We = 12 \ldots 20$,
- umbrella break-up for $We = 20 \ldots 35$,
- mixed break-up (of bag and umbrella type) for $We = 35 \ldots 70$, and
- ligament break-up for $We > 70$.

Below the limit of $We_{\mathrm{crit}} = 12$ a droplet may still be deformed, but remain stable as a single droplet. By observing the region of melt ligaments secondary break-up with high speed imaging, it was found, especially for large particles with a diameter above 200 µm, that the particles typically are nonspherical, but having an irregular shape. These droplets may still be in the state of deformation or spheroidization on break-up.

The break-up modes classification of Pilch and Erdmann [59] includes a description of break-up mechanism for high Weber numbers above $We > 350$. For very small Weber numbers, $We < 12$, a droplet may still break up into two equally sized droplets under certain conditions. According to this classification, Samenfink et al. [60] distinguished basically three mechanisms:

- bag break-up for $We = 12 \ldots 20$,
- multimode break-up for $We = 20 \ldots 70$, and
- shear break-up for $We > 70$.

For liquids of various viscosities, Hsiang and Faeth [61] derived an empirical correlation between the Sauter mean diameter and the Ohnesorge and Weber numbers. The initial droplet diameter before break-up is d_{part}.

$$\frac{d_{32}}{d_{\mathrm{part}}} = 1.5\,\mathrm{Oh}^{0.2}\,\mathrm{We}_{\mathrm{gas}}^{-0.25} \tag{5}$$

Typical values of liquids are in between $Oh^{0.2}$ $We_{\mathrm{gas}}{}^{-0.25} = 0.05 \ldots 0.4$, where metallic melts are also in this range. In case of shear break-up, the resulting droplets with new diameter d are root normal distributed as:

$$f(d) = \frac{\sqrt{\frac{d}{d_{50.3}}}}{2\sqrt{2\pi}\sigma d} e^{-\frac{1}{2}\left[\frac{\sqrt{\frac{d}{d_{50.3}}}-\mu}{\sigma}\right]^2}, \quad \text{where } \mu=1,\ \sigma=0.238 \tag{6}$$

$$d_{50,3} = 1.2\, d_{32} \tag{7}$$

This distribution has been applied by Schmehl [62] also for the bag- and multimode break-up regimes.

The Taylor analogy break-up model (TAB) [63] describes the resulting droplet diameter after secondary disintegration, where the Sauter mean diameter d_{32} for very low and high Weber numbers results in:

$$\frac{d_{32}}{d_{\mathrm{part}}} = \frac{3}{7} \text{ for } We \to We_{\mathrm{crit}} \text{ and } \frac{d_{32}}{d_{\mathrm{part}}} = \frac{12}{\rho_{\mathrm{gas}} u_{\mathrm{rel}}^2} \text{ for } We \to \infty \tag{8}$$

While the empirical model is applied in this work, the lower limit can be used at the critical point.

The dynamics of the break-up process also depend on the aerodynamic Weber number. In a simplified model, the molten droplet deforms to a disk and then breaks up into new droplets. A characteristic break-up time t^* is defined as:

$$t^* = \frac{d_{\mathrm{part}}}{u_{\mathrm{rel}}} \sqrt{\frac{\rho_{\mathrm{liquid}}}{\rho_{\mathrm{gas}}}}. \tag{9}$$

The deformation time is $t_{\mathrm{def}} = 1.6\, t^*$ due to [61]. The time needed for break-up is t_{breakup}

$= k \cdot t^*$, whereas k is an empirical constant depending on the break-up regime. The break-up time has been measured or estimated [63 – 68]. Compared to the break-up time, the spheroidization time is small, $t_{\text{spheroidisation}} = 0.88\ \mu_{\text{liquid}}\ d_{\text{part}}/\sigma << t_{\text{breakup}}$. For example, a molten steel droplet of $d_{\text{part}} = 100$ µm needs about 250 ns to spheroidize, which is many times faster than any possible break-up time. So it is assumed that new particles immediately after break-up are spherical.

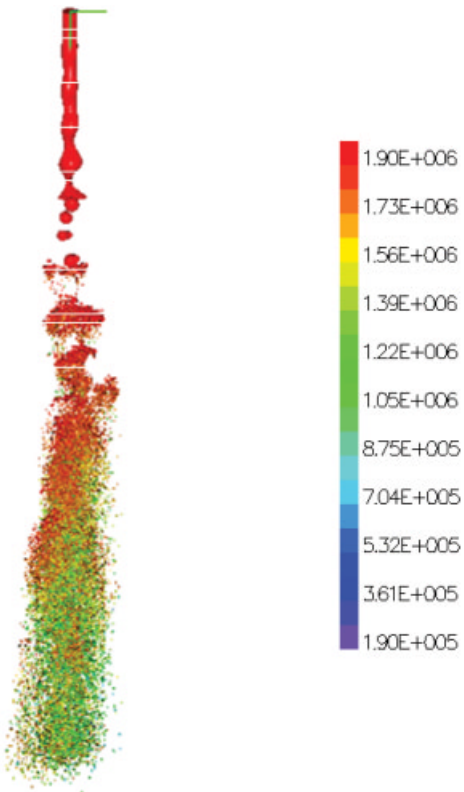

Figure 11. Instantaneous particle distribution for a spray atomized at 0.4 MPa, droplets coloring by specific enthalpy [55]

Figure 11 shows a typical instantaneous spatial distribution of the droplets in the spray. In the initial region of droplet production, a tendency to clustering of droplets is to be seen. Due to the sequential fragmentation of larger melt fragments that have been separated from the liquid jet, droplets are segregated in droplets clusters first. Further downstream these clusters are mixed to a more homogeneous droplet distribution mainly due to turbulence effects. The individual liquid elements and droplets are coloured by means of its specific enthalpy, where the cooling rate of droplets is illustrated.

The basic influence of the increased GMR and hence the nozzle pressure for a steel jet at 150 °C superheat atomized with nitrogen is shown in Figure 12. The resulting particle distribution is shifted from big to smaller particles for increasing nozzle distance.

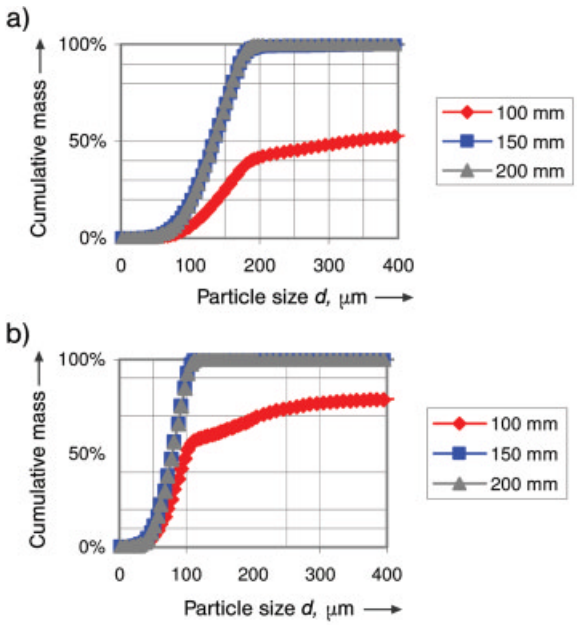

Figure 12. Size distribution at the distance x of the gas nozzle of an atomized steel jet, relative nozzle pressure $p_{\text{gas}} = 0.25$ MPa (A) and $p_{\text{gas}} = 0.4$ MPa (B) [55]

3. Spraying

The physical behavior of the spray cone is the key process for the development of a high-quality sprayed product [69, 70]. In the spray, a complex two-phase flow situation containing gas and particles/droplets occurs. The high-speed gas loses its momentum by accelerating the droplets. Due to the heat transfer from the droplets, the gas undergoes intense heating in the zone close to the atomizer. The droplets are accelerated and partly solidify during their flight. The mixture of different stages of droplet solid-

ification is believed to give the optimum conditions for producing a homogeneous preform. The largest droplets will reach the point of impingement in the liquid state. Due to their high mass they are still in the process of solidification. The medium-sized droplets are in a semisolid or mushy state, while the smallest particles are already fully solidified when they hit the deposit or substrate.

Measurement of droplet size and velocity distribution within molten metal sprays can be found in [71 – 74]. Continuous information on droplet behavior in the spray is obtained by numerical simulations [37]. Here, a model of the complete dispersed two-phase flow situation including momentum and heat coupling, is used to describe and analyze the complex spray cone behavior. Euler – Lagrange [75 – 77] or Euler – Euler modeling is used [78, 79] for multiphase flow analysis. In both types of multiphase flow model, the gaseous phase is described as a continuum (Eulerian approach) by means of conservation equations for momentum, mass, and energy. In an Euler – Euler model the dispersed phase is also described as a quasi-continuum, based on the same conservation equations, averaging the particle properties over a finite volume. In an Euler – Lagrange model of the dispersed phase, individual particles (or particle parcels) are tracked through the flow field. Depending on the material used and alloy composition, different solidification models have to be included in the simulation.

A high-speed-pyrometry system may be used for monitoring of the spray forming process [80]. The use of high-speed pyrometry in a two-phase flow under spray forming conditions enables the determination of temperature, velocity and size of in-flight particles without any influence on the measuring environment and on the spray itself. This measurement device consists of three main components as illustrated in Figure 13: a) a sensing head located near the spray cone collecting the thermal radiation emitted from the hot particles, b) the detection cabinet containing the optical components and photodetectors, and c) a computer equipped with the required digitizing and computing boards. The sensing head is aimed perpendicular to the spray cone collecting the light emitted by in flight particles. The collected light is imaged on the end of the optical fiber bundle. The core fiber is covered by a two-slit photomask. Each slit width and the distance between the slits is designed for a specific measurement range. The detection system for the central fiber is a high-speed, two-color pyrometer where the light is focused through a dichronic mirror onto two photodetectors. The signals are statistically compiled, displayed, and saved. Light from a particle traveling in the sensors focal plane enters the central fiber only when at least part of its image is located within one of the masks two slits. The characteristic signature of a particle crossing the sensors focal point is a two-peak signal representing 100 %, 0 %, and again 100 % of its radiation within the corresponding detector bandwidth. The shape of

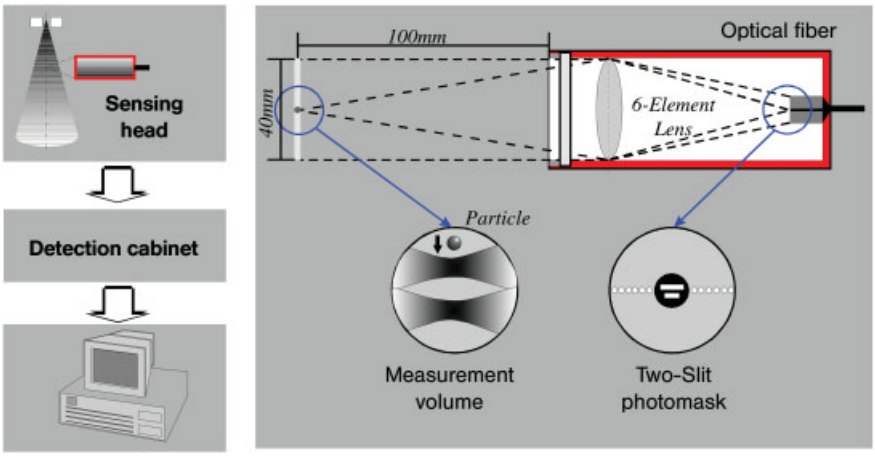

Figure 13. System components of an in-flight high-speed particle pyrometer [73]

the measured two-peak signal depends on and changes with the particle velocity, temperature, diameter, and trajectory. The particle velocity, v_p, is computed from measuring the time interval between the two peaks. The particle temperature, T_p, is obtained from the ratio of signal intensities from each detector according to Planck's law and assuming that emissivity of particle surface is the same at both wavelengths ($\varepsilon(\lambda_1) = \varepsilon(\lambda_2)$):

$$T_p = \frac{c_2\,(\lambda_1 - \lambda_2)}{\lambda_1 \cdot \lambda_2} \cdot \left[\frac{1}{\ln R + 5 \cdot \ln \frac{\lambda_1}{\lambda_2}} \right] \tag{10}$$

where c_2 is the second radiation constant and R is defined as the energy ratio of $E(\lambda_1)/E(\lambda_2)$. Both energies are calculated through the time integral of the two-peak signal normalized by the velocity. Using Planck's law by taking into consideration that melted particles are spherical or at least very close to be, one can derive the diameter of a particle d_p from

$$d_p = \sqrt{\frac{E\,(\lambda_i)}{c_3 \cdot \varepsilon\,(\lambda_i)}} = \sqrt{\frac{E\,(\lambda_i)}{d_c}} \tag{11}$$

where i represents 1 or 2 and c_3 is the third radiation constant. Since $\varepsilon(\lambda_i)$ is extremely difficult to define, the calculation system includes a coefficient d_c, which includes both c_3 and $\varepsilon(\lambda_i)$. For a new material with a thermal history emissivity $\varepsilon(\lambda_i)$, and thus d_c, is unknown at the beginning of the experiment. In order to obtain precise d_p values, one must follow, for each different material, a calibration procedure.

Spray forming of high carbon steel (100Cr6, containing 1.35 to 1.6 % chromium) exhibit the influence of different spraying conditions on the particles thermal and kinetic state. On the spray center line the axial distance is varied in three steps as $z = 150$ mm, $z = 200$ mm, and $z = 325$ mm from the nozzle. The melt superheat of $\Delta T = 150\,°C$ and the atomizing gas pressure of $p_g = 2.5$ bar were constant. Figure 14 demonstrates the measured particle temperature versus the cumulative number distribution (counts in %) with increasing axial distance. It is important to point out that even in the nearest position of the spray cone ($z = 150$ mm) most of the particles in spray forming process have a temperature beneath the state of phase change (for the used high-carbon steel the liquidus temperature is $T_l = 1450\,°C$, and the solidus temperature is $T_s = 1300\,°C$, the temperature range of solidification will be identified in the temperature plots by a grey bar). As to be seen from Figure 15, the particle velocity increases slightly with increasing measurement position indicating that on average, the droplets were continuing to accelerate at these axial distances. The increase of particle diameter illustrated in Figure 16 occurs because of the higher radial dispersion of smaller particles in comparison to that of larger particles with more mass inertia. The general flow direction and turbulence of the gas carries them away from the centre to outer parts of the spray cone with increasing distance from the atomization nozzle.

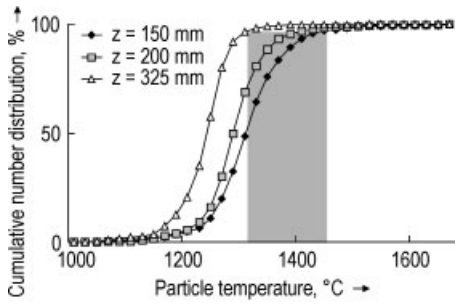

Figure 14. Change in particle temperature with variation of the measurements position (grey area = solidification temperature range) [73]

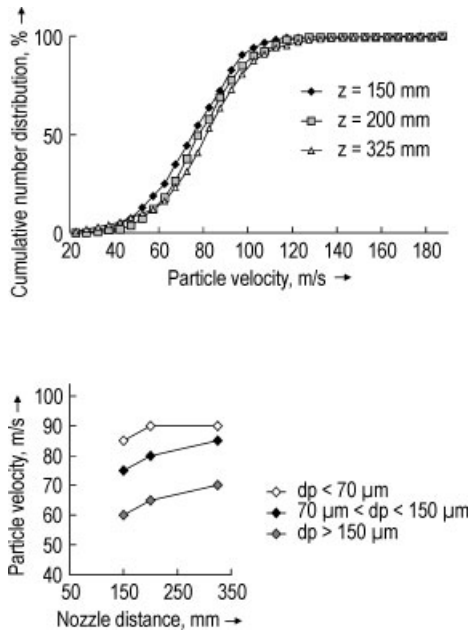

Figure 15. Change in particle velocity through variation of the measurements position [73]

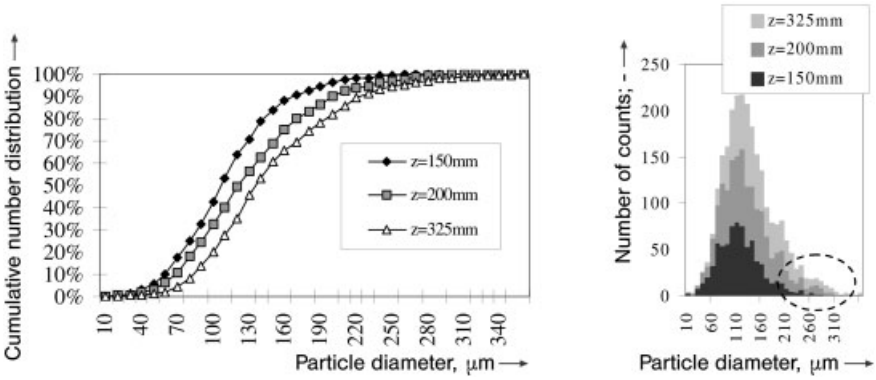

Figure 16. Change in particle diameter with variation of the measurements position [73]

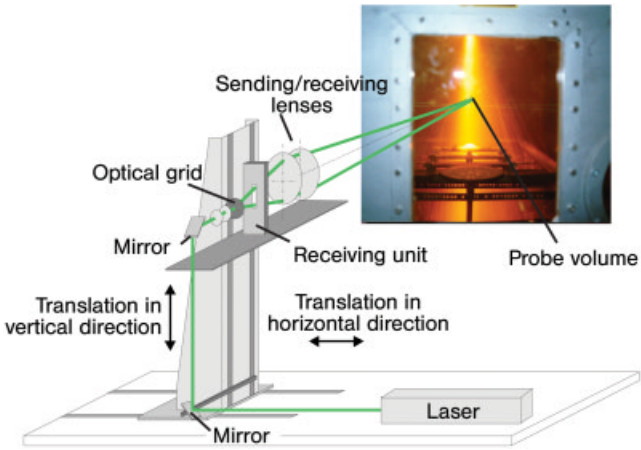

Figure 17. Scheme of backscatter PDA optical setup for spray forming application [74]

Phase-Doppler-Anemometry (PDA) is another suitable process measurement technique to detect size and velocity of spherical particles with smooth surfaces simultaneously in multiphase flows. However, sprayed droplets and particles within liquid metal atomization do not fulfill these requirements. Due to high cooling rates in flight some particles are partially or completely solidified exhibiting a rough surface due to dendritic solidification. This surface roughness causes noisy and weak signal response within light scattering. In order to overcome this problem, as well as to tackle difficulties resulting from the hostile environment inside the spray chamber during spray forming, an adaptation and modification of the common PDA technique is necessary. A detailed description of conventional PDA techniques is given in [81 – 83]. Specifically the signal processing had to be modified because of the poor signal quality [84, 85]. The optical part of the PDA is realized in direct backscattering direction, as illustrated in Figure 17. Consequently, a compact moveable PDA is used where only a single window in the spray chamber for optical access is needed. The ability to move the measurement volume in vertical as well as in horizontal direction during one spray forming run allows the investigation of different measuring locations in the spray.

Using this modified PDA, the spray cone resulting from the atomization of the liquid metal jet can be investigated with regard to different process parameters, e.g., gas flow, melt temperature, and measurement position in the spray cone [86].

Using this modified PDA, not only particle size and velocity in the two-phase flow can be measured but also the local mass flux can be determined. The PDA results have been verified by comparing this data to results received via sieve

analysis of collected particles as illustrated in Figure 18. Measurements carried out with usage of a scanning atomizer, resulting in a movement of the spray cone relative to the (fixed) interference volume, demonstrates the sensitivity of the employed measurement technique [74, 87]. In order to generate definite deposit geometries by spray forming, in most cases the movement of the atomizer is necessary. Varying the position of the measurement volume relating to the substrate/deposit, the mass distribution in dependence on the local position above the deposit can be estimated.

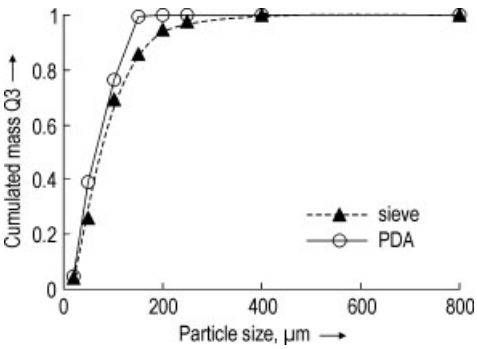

Figure 18. The cumulated mass versus particle size obtained by using PDA and sieve analysis, respectively [87]

By extending the PDA technique to an on-line measurement system, even process analysis very close to real time becomes possible and the PDA can be used to monitor the spray forming process [88].

For this, the Fourier based signal processing has to be realized via hardware, i.e., a special PDA signal processor has been developed, the use of which allows the data evaluation in only a couple of seconds. The signal flow is as follows:

1) The analog PDA signal is converted to a digital one,
2) the signal is computed in a digital signal processor (DSP) considering the introduced Fourier analysis,
3) the resulting diameter and velocity distributions can be used for both monitoring the atomization process and controlling the process,
4) for controlling the process an adjustment signal is created by comparing the actual value of the mean particle diameter (or the mean particle velocity) to the corresponding set value,
5) the digital adjustment signal is converted to an analog signal, the quantity of which is influencing the valve for the atomization gas pressure.

In its most simplified form, a gas control loop consists of the PDA measurement unit, the regulating valve and the control unit allowing the automatic adjustment of the mean particle velocity or diameter by varying the atomization gas pressure. Analyzing the atomization process taking place during spray forming the significant influence of atomization gas pressure on particle properties can be shown [89, 92]. Since particle diameter and velocity are essential to heat and mass transfer, among others these process parameters are defining the thermal conditions of the impacting particles and with it the properties of the spray-formed material like microstructure, grain size, and chemical homogeneity. Thus, the aim is to monitor and to control the spray forming process on the basis of the mean particle diameter and velocity, respectively. In order to investigate the control loop and the on-line PDA, respectively, in this section sharp variations in gas pressure are realized, purpose of which is not to produce a high-graded deposit but to examine the temporal behavior of the atomization process and in addition that of the on-line PDA unit.

Figure 19 summarizes the on-line measured mean particle velocity and the corresponding atomization pressure versus duration of one spray forming run. In case of particle velocity the actual as well as the desired value are shown, i.e., using the closed control loop a definite particle velocity $v_{1,0,\mathrm{set}}$ is proposed and the required adjustment of the current particle velocity $v_{1,0,\mathrm{act}}$ by varying the atomization gas pressure p_{g} is observed. The graphs agree with the well-known correlation of particle velocity and gas pressure. Therefore, with an increasing gas speed the immediate increase of particle velocity can be made out using PDA. By contrast, some more time is needed to adjust the actual particle velocity. In Figure 19 time intervals of $\Delta t = 27$ s and $\Delta t = 35$ s are needed to realize the necessary adaptation of the current particle velocity to the set value. Nevertheless, time consumption for ad-

justment is of that scale that influencing control throughout a current process is realizable.

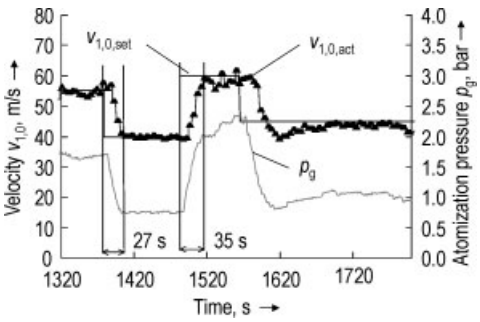

Figure 19. Automatic control of the particle velocity [74]

The second PDA measurable particle property is the diameter. Within the scope of verifying the ability of diameter control the set-value of the mean particle diameter $d_{1,0,\mathrm{set}}$ is placed several times far from the actual value $d_{1,0,\mathrm{act}}$ (corresponding to Figure 20) bringing modifications of the gas pressure as large as possible. It can be seen that the corresponding mean diameter of the metal droplets is following in the expected way, even if only a slight change in the average diameter is noticeable and a straggling conditional on fluctuations of both the measurement technique and the process itself are unavoidable. Using PDA as an on-line measurement system and adapting the on-line PDA to a closed control loop for the atomization gas enables the measurement and influencing of the metal disintegration process. Hence, process control very close to real time becomes possible. Subsequently, well-aimed variations in the atomization result can be created and measured by varying the atomization gas pressure within the scope of automatic control operations.

As an example of a modeling result for the spray cone behaviour, Figures 21 and 22 show the behavior of different droplet classes on the centerline of a spray of molten steel particles atomized with nitrogen [75]. The boundary conditions include an atomizer pre-pressure of p_2 = 3 bar abs., a gas to metal flow rate (GMR) of 1.25 and a droplet size distribution of log-normal type with a mass median droplet size of 133 μm. The axial velocity distribution on the spray cone centerline for gas and particles with different diameters is illustrated in Figure 21. The gas velocity far upstream of the atomization area is unaffected by the particles. Close to and in the atomization region, the gas velocity decreases due to the momentum transfer for acceleration of the droplets. Afterwards, the gas velocity increases again because of intense heating of the gas and the resulting thermal expansion and also due to the entrainmaint of momentum from the edge of the spray cone. The maximum gas velocity occurs at a distance of ca. 330 mm from the nozzle. Afterwards, the gas velocity decreases due to spreading of the spray cone. Close to the point of spray cone impingement onto the deposit/substrate, the axial gas velocity rapidly slows to zero. Particles are accelerated in the atomization region and further downstream. In the main region of the spray cone, particle velocities of the larger particles remain almost constant.

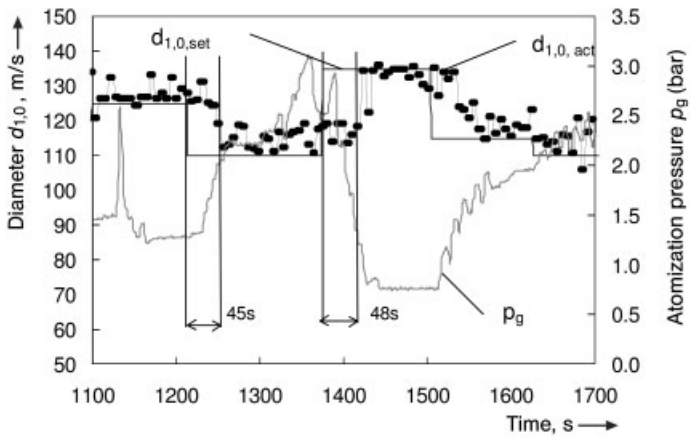

Figure 20. Automatic control of the particle diameter [74]

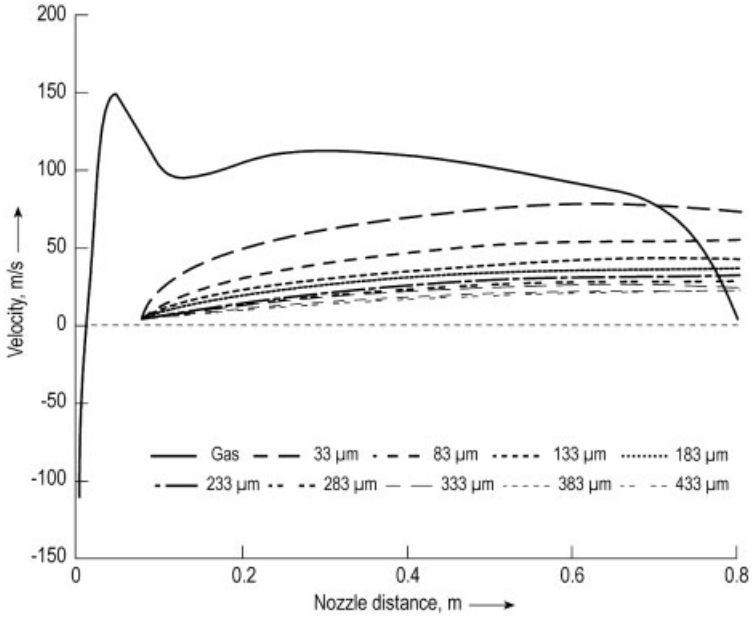

Figure 21. Gas and particle velocities on the spray cone centerline [75]

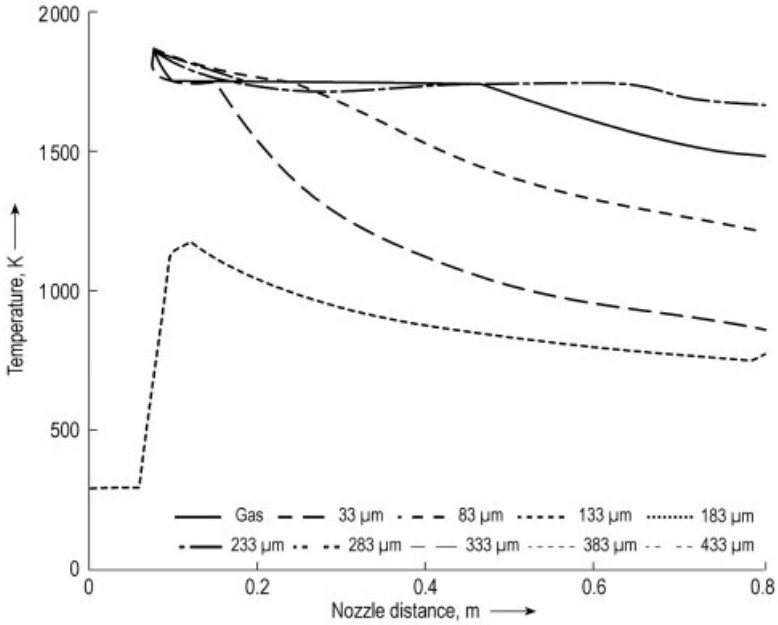

Figure 22. Gas and particle temperature on the spray cone centerline [75]

The gas temperature in the atomization region reaches values of up to 1200 K (Fig. 22) [75]. Further downstream, the local gas temperature on the centerline decreases, but heat transfer from the particles still occurs, due to the spreading of the spray and mixing with colder ambient gas. Particles are assumed to be preheated to 125 K above the solidification temperature at the point of atomization. After cooling to the solidification temperature, the phase change occurs inside the particles and the latent heat of the particles is released, whereby the particle tem-

perature remains constant. When the latent heat content of the droplets has been removed, the particles cool further. From these calculations it can be seen, that particles bigger than 210 µm are still in the state of phase change when they impact the substrate/deposit.

The mass flux distribution in the spray cone is directly related to the shape of the sprayed deposit. For varying deposit shapes, either the atomizer and/or the substrate can be moved. For a stationary atomizer, the radial profile of the mass flux distribution in the spray cone can be described as an exponential function by:

$$\dot{m}_{\mathrm{spray}}(r) = \dot{m}_{\mathrm{max}} \exp\left(\ln 0.5 \left|\frac{r}{r_{0.5}}\right|^{k_1}\right) \tag{12}$$

The maximum mass flux is always found on the spray cone centerline and $r_{0.5}$ is the half-width of the mass flux profile (see Fig. 23). The exponential factor is correlated as k_1 = 1.3 [90, 91] and k_1 = 1.124 [92] for steel atomization. The half-width of the mass flux profile is proportional to the maximum mass flux value:

$$\dot{m}_{\mathrm{max}}(x) \sim \frac{1}{r_{0.5}^2(x)} \tag{13}$$

For calculating the mass flux distribution at each position within the spray cone, a single calibrating measurement at one point on the centerline in combination with the mass flux correlation is sufficient.

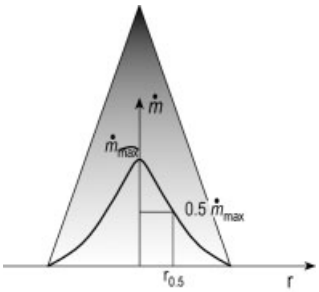

Figure 23. Principle of mass flux distribution profile

In a similar way, the thermal energy flux (enthalpy flux) within the spray may be detected and described [93, 94].

4. Deposition

Droplet deformation during droplet/particle impingement onto the substrate/deposit is described in [95 – 98]. A possible reason for micropore formation in spray forming and thermal spray processes based on the combined deformation and solidification of droplets is given in [99, 100]. Here it is shown that ejection of the liquid material over a rapidly solidified base layer of the splat may result in a gap or pore.

The impinging droplets transport mass, momentum and thermal energy into the top layer of the deposit [93]. The compacted mass is not identical to the theoretical impinging mass from the spray because

- Smaller droplets of low inertia follow the gas streamlines and are deflected
- Droplets may bounce off or splash and secondary droplets may not be incorporated in the deposit
- Particles with high momentum may tear off parts of the already compacted mushy top layer

These particles contribute to the overspray. The difference between the mass flux value in the spray cone and the compacted mass is described by the deposition efficiency:

$$k(\vec{x},t) = \frac{\dot{m}_{\mathrm{compac}}(\vec{x},t)}{\dot{m}_{\mathrm{spray}}(\vec{x})} \leq 1 \tag{14}$$

The deposition efficiency is a function in space and time and is mainly proportional to:

- The thermal state of the impingement surface
- The momentum, temperature and state of solidification of the impinging droplets

The deposition efficiency for spray forming of Gaussian shaped deposits (produced with a stationary atomizer, spraying on a stationary cylindrical substrate) from steel is given by [92]:

$$k(T_{\mathrm{p}}, f_{\mathrm{L}}, I_{\mathrm{p}}) = C_4 \left[C_1 + C_2 \left(\frac{T_{\mathrm{p}}}{T_{\mathrm{L}}}\right)^{C_3}\right] f_{\mathrm{L}}^{C_5} I_{\mathrm{p}}^{C_6} \tag{15}$$

where the coefficients C_1 to C_6 are:

C_1	0.422	C_4	0.489
C_2	0.800	C_5	0.249
C_3	30.48	C_6	– 0.062

The main influencing parameter for the deposition efficiency is the temperature of the droplets at impact (here normalized with respect to the liquidus temperature).

During spraying, the deposit grows continuously and releases its heat by conduction to the substrate and convection and radiation to the surrounding gas.

If the particles are not fully solidified prior to impact on the substrate, which is a fundamental condition for the compaction of particles to a homogeneous preform, a solidification front must exist at some point in the deposit. To calculate the spatial and temporal temperature distribution in the deposit and substrate, numerical methods have been used [6, 101 – 104], because of the difficulty of temperature measurement in the deposit. These models consider the deposition rate, but are limited to one- or two-dimensional configurations. In an extended approach, a more accurate and more general formulation of the problem is given and solved in [105, 106].

By numerical modeling, a description of the cooling process of metals during flight, deposition and further cooling periods in spray forming (which are of major importance for the understanding, control and optimization of the process) can be obtained and analyzed. The quality of spray-formed products in terms of homogeneity and improved material properties is directly controlled by the cooling rate of particles in the spray cone and the further cooling behavior of the resulting preform. The solidification and cooling behavior of the material directly depends on operational parameters.

A result of a thermal calculation in the deposit and substrate during spray forming of a Gaussian shaped steel deposit is shown in Figures 24 and 25 [105, 107]. The transient contour of a growing deposit and the calculated surface temperature is shown in Figure 24 after 15 and 30 s spray time. Then the spray process ends, and the cooling period begins. Therefore, the deposit shape does not change any further, but the surface temperature decreases due to cooling from the gas flow. In Figure 25 the temperature distribution in the central plane inside the deposit and underlying substrate is illustrated at the same time intervals as in Figure 24. The calculated temperature behavior shows a temperature difference immediately at the interface between deposit and substrate. Here, due to a residual oxide layer and pore formation in the first compacted material layers, a heat transfer resistance results and decreases the heat transfer from the deposit into the substrate [4]. After 60 s total time, a hot-spot area inside the deposit can be identified. The continued cooling of the gas flow results in a colder shell around a hot core. This area is extremely problematic with respect to grain coarsening and (for spray forming of solid preforms) hot shortness. Temperature simulations of the heat balance in the deposit and substrate can reveal possibilities for avoiding hot spots.

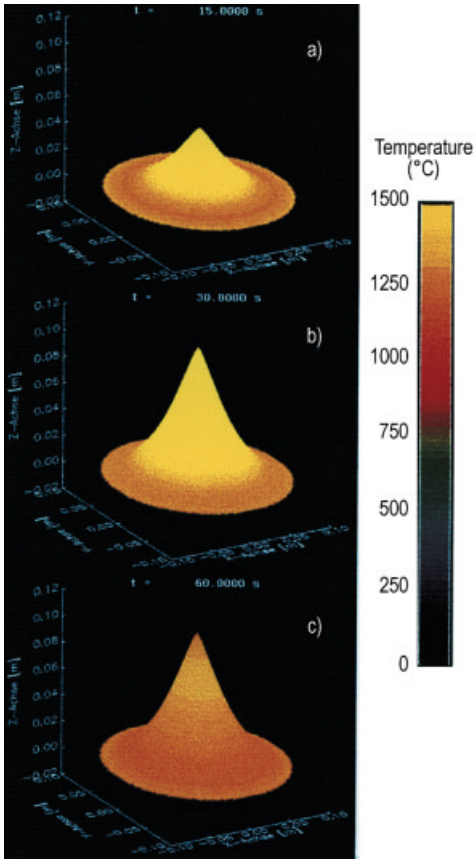

Figure 24. Surface temperature of spray-formed Gaussian-shaped deposit

Problems in spray forming may arise from remaining gas pores in the preform. These may occur, e.g., because of gas entrapment (too hot spray conditions) or unsufficient liquid content in the spray (too cold spray conditions) The possible formation of macro- and micropores in spray-formed deposits has been investigated within modeling of the spray solidifying on an arbitrary surface within a geometry simulation by Djuric et al. [108, 109]. As also has been

pointed out, e.g., by Cai and Lavernia [110], the presence of some remaining porosity is inevitable in spray-formed materials. Typically zero or low porosity (< 1 %) is found in the largest region in the core of spray-formed preforms, but the porosity level may increase up to 10 % within the surface area of the preform. From a quality point of view, porosity in spray deposited materials yields two types of problems. First, material properties may degrade because of the presence of pores. Second, secondary working (such as extrusion, rolling, forging, and HIPing) is necessary to achieve full density, which in turn limits the suitability of spray forming as a near-net shape manufacturing process. Consequently, research and modeling efforts have been initiated in spray forming to better understand the formation mechanisms of porosity formation and determine ways to decrease the porosity level either by adjusting the process parameters or by exploring alternative approaches.

The total porosity in an as-deposited material in spray forming can be described as the sum of:

- porosity from gas entrapment,
- interstitial porosity,
- porosity associated with shrinkage during solidification.

Among these mechanisms, interstitial porosity has been reported to play the most important role in porosity formation during spray forming, particularly when spraying thin geometries like bands and strips or tube clads. Considering that the formation of interstitial porosity is due to the lack of liquid to fill the overlapped particle interstices, whereas solidification porosity is significant when too much liquid is present in the spray upon impingement, interstitial porosity and solidification porosity may be treated as two mutually exclusive occurrences. Consequently, in a modeling approach, when interstitial porosity is predominant, solidification porosity is ignored (and vice versa). In a first modeling approach [110], the gas entrapment porosity has been neglected. The model starts to describe the consolidation mechanism in two stages. In the first stage, solidified droplets impinge on the deposition surface and form a random-dense particle packing. Within the second stage, liquid droplets first impinge on the resultant particle packing and finally solidify. If the volume of liquid, once solidified, is smaller than that of voids in the particle packing, porosity is assumed to be dominated by interstitial porosity, whereas the solidification shrinkage is neglected. On the other hand, if the volume of liquid, once solidified, is more than that corresponding to voids, porosity is assumed to be controlled by solidification of the liquid left and interstitial porosity is neglected – solidification porosity is dominating. For this model, the particle size distribution in the spray has been assumed as of log-normal type, the median particle diameter has been adopted from Lubanska's formula (Eq. 1), and the standard deviation of the drop size distribution has been taken from empirical metal atomization data [111].

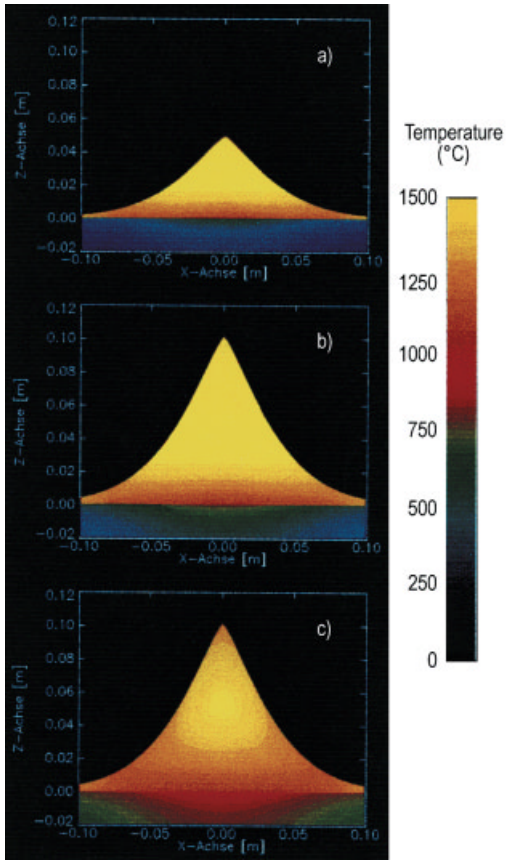

Figure 25. Temperature inside Gaussian-shaped deposit and substrate during spray forming

The porosity model has been applied to spray forming of an Al-4-wt% Cu alloy, atomized

by nitrogen. The main process parameters have been varied as the atomization gas pressure in a range from 1 to 10 MPa, the melt flow rate between 0.001 and 0.02 kg/s, and the melt superheat from 50 to 250 K. The temperature of the atomization gas has been assumed as constant. A typical result of the simulation is shown in Figure 26. Here, the variation of porosity with deposition distance for an atomization gas pressure of 1.2 MPa, a melt flow rate of 0.01 kg/s and varying melt superheats is illustrated. The porosity level exhibits a V-shaped variation with deposition distance, reflecting both types of porosity reasons, the solidification and the interstitial porosity. For small deposition distances the spray is too hot and solidification porosity results. As deposition distance increases, porosity decreases continuously to such a point where minimum porosity is achieved. Beyond this point, porosity increases almost linearly with increasing deposition distance. Here, the spray conditions become too cold and the porosity is caused be interstitial effects. As an example, for a melt superheat of 100 K, porosity is 5 % at deposition distance of 0.1 m. The point corresponding to a minimum porosity in this case is at a deposition distance of 0.36 m. For increasing melt superheat, the minimum porosity is achieved at longer deposition distances.

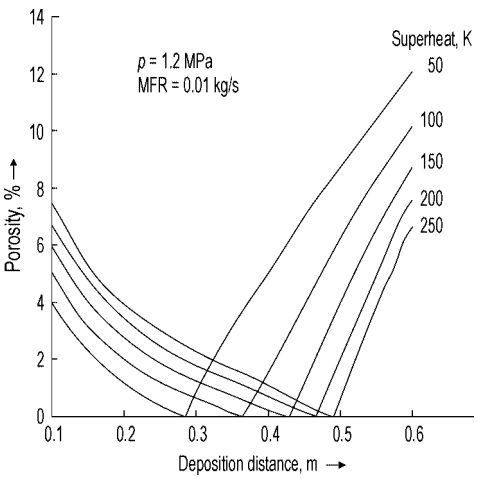

Figure 26. Porosity as a function of deposition distance for an atomization gas pressure of 1.2 MPa and a melt flow rate of 0.01 kg/s for Al-4Cu [110]

The variation of other process parameters, as the atomization gas pressure and also the melt flow rate, result in principle in similar V-shaped porosity curves as for varying the deposition distance. Based on these results, optimum combinations of processing parameters during spray forming may be identified. A porosity coefficient may be introduced [110] to identify dominant mechanisms in the different porosity formation regimes. In addition to porosity, other process constrains, such as microstructure as well as process economy, have to be taken into consideration for finding a set of suitable spray forming process conditions.

The model study on porosity formation in spray forming [110] has only the spray conditions taken into account. For further model developments, as main influencing parameter for porosity, also the state of the deposit surface has to be taken into account. This may be achieved mainly in terms of deposit of the surface temperature, which severely affects the porosity as well as the particle sticking efficiency [112].

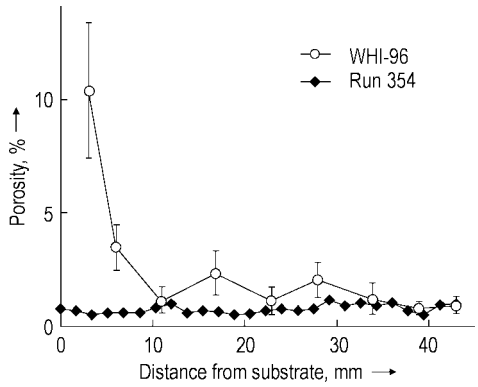

Figure 27. Local porosity in as-sprayed IN718 rings on cold substrate (marked as Whi-96) and this study (run 354), atomizer gas is argon [113]

The porosity at the bottom of spray-formed products may be successfully decreased by preheating the substrate surface. This has been demonstrated in [113] for spray forming of rings from superalloys IN718 and U720. Within this process the melt was superheated (app. 150 K above liquidus temperature) in the crucible. Subsequently, the melt (150 kg) was pulled into the preheated tundish and left the nozzle where it is atomized. A rotating cylindrical shaped sub-

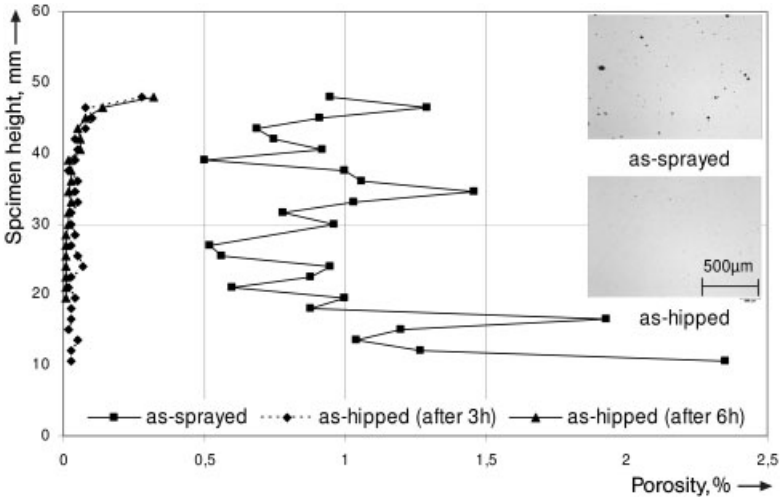

Figure 28. Porosity of spray-formed IN718 before and after HIP [113]

strate was preheated to app. 1150 °C by an induction coil from the inside of the rotating substrate (rotation frequency 1.2 Hz). By using the preheated substrate the base porosity of the deposited ring structure has been drastically decreased as illustrated in Figure 27 in a comparison to a sprayed ring using a cold substrate. After subsequent HIPing of the sprayed rings, the overall material porosity may be decreased below 0.1 % in most parts of the deposit (Fig. 28).

5. Integral Modeling

Modeling and simulation of the spray forming process is an essential tool for process understanding and development [37, 114, 115]. Integral spray forming models especially for the thermal behaviour in spray forming have been derived by Bergmann [116], Minisandram et al. [117], and Pedersen [118 – 120]. Connection of various submodels is performed for transient-temperature material behaviour, from melt superheating in the tundish to room temperature in the preform via cooling and solidification. By inclusion of a tundish melt flow and thermal model in a first stage, employment of the local separation method for spray properties averaged temperatures and solid fractions for the particle mass on the centre line of the spray is done in a second stage. The time axis is shown in terms of the mass averaged mean residence time of the particle mass. By combination of these data with the calculated temporal cooling and solidification distributions of a fixed volume element inside the deposit, one yields the mean thermal history of the material at a specified location in the deposit.

The transient thermal and solidification distributions are shown in Figures 29 and 30. Results are illustrated separately for the three different modeling areas (with different time scales): a) melt flow in the tundish, b) particle cooling in the spray, c) growth and cooling of the deposit. The process conditions used are for a steel spray forming experiment for production of a Gaussian shaped deposit.

From top of the tundish (upper melt surface) to the exit of the tundish in the atomization nozzle, the melt is cooled by approx 20 K (for an initial melt temperature of $T_L = 1933$ K). Solid fraction of the melt does not change as the melt stays fully liquid ($f_s = 0$). In total, the mean residence time of the melt in the tundish is 51.6 s for an exit diameter of $d_{exit} = 4$ mm of the tundish and a surface height of $h = 0.25$ m of the melt in the tundish (determining the melt mass flow rate). A velocity profile within the melt flow develops in the tundish with highest velocity values on the centre line.

In the spray flow, mainly convective heat transfer to the gas cools the melt droplets. Until the spray impingement onto the substrate ($z =$

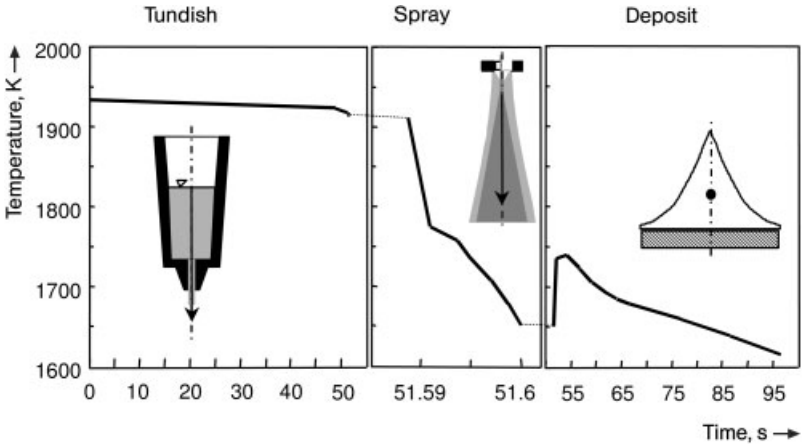

Figure 29. Integral thermal spray forming model: Temperature evolution from tundish to deposition [116]

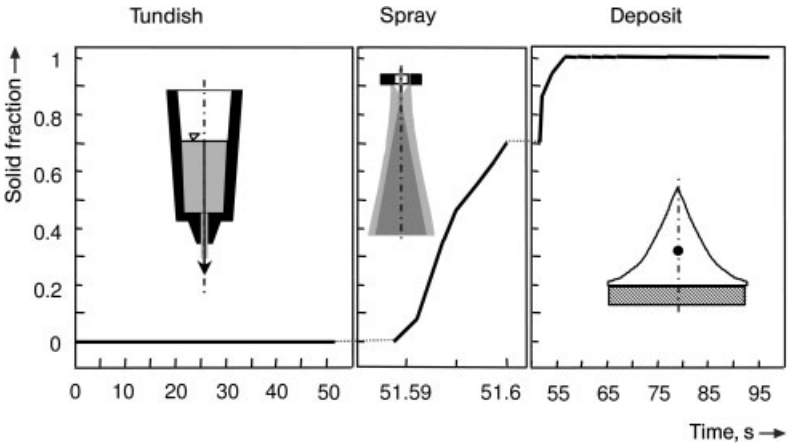

Figure 30. Integral thermal spray forming model: Solid fraction evolution from tundish to deposition [116]

0.6 m; $s = 0.47$ m) the mean residence or flight time of the particles in the spray is 11 ms. In this time the melt material is cooled down by $\Delta T =$ 260 K ($T_m = 1650$ K) corresponding to a mean cooling rate of 2.4 10^4 K/s. Hereby the mean solid fraction in the spray increases to $f_s = 0.71$.

Because of thermal diffusion and the release of latent heat inside the deposit, the material temperature directly after compaction at first increases by further increasing the local solid fraction. For the discussed parameter setting the temperature increase is approx 90 K. After complete solidification of the volume element under investigation ($f_s = 1$) the temperature will decrease again by a cooling velocity of approx. 2 to 3 K/s.

6. References

1. A. R. E. Singer: *Koll. SFB 372*, vol. **1**, Universität Bremen 1996, pp. 123 – 140.
2. Patent No. P 25 37 103.3 - 24, 20.8.75 (R. G. Brooks, S. Green).
3. S. Annavarapu, D. Apelian, A. Lawley, *Metall. Trans.* **A 19** (1988) 3077 – 3086.
4. E. J. Lavernia, Y. Wu: *Spray Atomization and Deposition*, J. Wiley & Sons, Chichester 1996.
5. A. G. Leatham, A. Lawley, *Int. J. Powder Metall.* **29** (1993) 4, 321 – 329.
6. P. Mathur, D. Apelian, A. Lawley, *Acta metall.* **37** (1989) 2, 429 – 443.
7. K. Bauckhage, Gaithersburg *Proc. ICLASS-91*, 1991, pp. 33 – 48.

8. C. Y. A. Tsao, N. J. Grant, *Int. J. Powder Metallurgy* **30** (1994) 3, 323 – 333.
9. W. Reichelt, *Koll. SFB 372*, vol. **1**, Universität Bremen 1996, 189 – 198.
10. A. Lawley, Proc. SDMA 2000, Bremen, pp. 3–16.
11. K. Bauckhage, *HTM 52* (1997) 5, 319 – 331.
12. S. Markus, U. Fritsching, Proc. Spray Deposition and Melt Atomization SDMA 2003, Bremen.
13. S. Markus, C. Cui, U. Fritsching, *Mater. Sci. Eng. A* **383** (2004) 166–174.
14. J. V. Wood (ed.): Proc. 2nd Int. Conf. Spray Forming, Swansea 1993, Woodhead Publ. Ltd. Cambridge 1993.
15. J. V. Wood (ed.): Proc. 3rd Int. Conf. Spray Forming, Cardiff 1996, Osprey Metals Ltd. 1997.
16. J.V. Wood (ed.): Proc. 4th Int. Conf. Spray Forming, Baltimore, USA 1999, Osprey Metals Ltd. and Welsh Development Agency 1999.
17. K. Bauckhage, V. Uhlenwinkel (eds.): *Sprühkompaktieren — Sprayforming*, Kolloquium des SFB 372, vol. **1**, Universität Bremen 1996.
18. K. Bauckhage, V. Uhlenwinkel (eds.): *Sprühkompaktieren — Sprayforming*, Kolloquium des SFB 372, vol. **2**, Universität Bremen 1997.
19. K. Bauckhage, V. Uhlenwinkel (eds.): Sprühkompaktieren — Sprayforming, Kolloquium des SFB 372, **vol. 3**, Universität Bremen 1998.
20. K. Bauckhage,V. Uhlenwinkel (eds.): Sprühkompaktieren — Sprayforming, Kolloquium des SFB 372, **vol. 4**, Universitüt Bremen 1999.
21. K. Bauckhage,V. Uhlenwinkel (eds.): Sprühkompaktieren — Sprayforming, Kolloquium des SFB 372, **vol. 5**, Universität Bremen 2001.
22. K. Bauckhage,V. Uhlenwinkel (eds.): Sprühkompaktieren — Sprayforming, Kolloquium des SFB 372, **vol. 6**, Universität Bremen 2002.
23. K. Bauckhage,V. Uhlenwinkel (eds.): Sprühkompaktieren — Sprayforming, Kolloquium des SFB 372, **vol. 7**, Universität Bremen 2004.
24. K. Bauckhage, V. Uhlenwinkel, U. Fritsching (eds.): Proceedings of the Spray Deposition and Melt Atomization Conference SDMA 2000, Bremen, Univ. Bremen 2000.
25. K. Bauckhage, V. Uhlenwinkel, U. Fritsching (eds.): Proceedings of the 2nd Spray Deposition and Melt Atomization Conference SDMA 2003, Bremen, Univ. Bremen 2003.
26. H. R. Müller: *Koll. SFB 372*, vol. **1**, Universität Bremen 1996, pp. 33 – 56.
27. K. Hummert: *Koll. SFB 372*, vol. **1**, Universität Bremen 1996, pp. 199 – 215.
28. R. L. Kozarek, D. D. León, A. Mansour: *Koll. SFB 372*, vol. **1**, Universität Bremen 1996, pp. 141 – 160.
29. K. Wünnenberg: *Koll. SFB 372*, vol. **1**, Universität Bremen 1996, pp. 1 – 32.
30. C. Spiegelhauer, L. Shaw, J. Overgaard, G. Oaks: *Koll. SFB 372*, vol. **1**, Universität Bremen 1996, pp. 57 – 62.
31. C. Spiegelhauer, *Koll. SFB* **372**, vol. 5 Bremen 2001, pp. 63–68.
32. F. G. Müller, M. G. Benz, W. T. Carter Jr., R. M. Forbes, A. Leatham: *Koll. SFB 372*, vol. **1**, Universität Bremen 1996, pp. 169 – 188.
33. E. J. Lavernia: *Koll. SFB 372*, vol. **1**, Universität Bremen 1996, pp. 63 – 122.
34. T. Ebert, F. Moll, K. U. Krainer, *Proc. 3rd Int. Conf. Spray Forming*, Cardiff 1996, 177 – 185.
35. A. Lawley, P. Mathur, D. Apelian, A. Meystel, *Powder Metall.* **33** (1990) 109 – 111.
36. P. Ottosen, Ph. D. Thesis Tech. Univ. Denmark TM.93.27 1993.
37. U. Fritsching: *Spray Simulation: Modeling and Numerical Simulation of Sprayforming Metals*, Cambridge University Press, Cambridge, UK, 2004.
38. U. Fritsching, H. Liu, K. Bauckhage, *Proc. ICLASS-91*, Gaithersburg, M.D., USA, NIST SP813 (1991), pp. 491 – 498.
39. K. Bauckhage, V. Uhlenwinkel, *HTM* **51** (1996) 5, 289 – 297.
40. J. Tillwick, H. Dahl, V. Uhlenwinkel, K. Bauckhage, T. Wriedt, *Technisches Messen* **64** (1997) 12, 481 – 490.
41. K. Bauckhage, *Chem. Ing. Tech.* **64** (1992) 4, 322 – 332.
42. A. J. Yule, J. J. Dunkley: *Atomization of Melts*, Clarendon Press, Oxford 1994.
43. K. Bauckhage, U. Fritsching in: K. P. Cooper, I. E. Anderson, S. D. Ridder, F. S. Biancanello (eds.): *Liquid Metal Atomization: Fundamentals and Practice*, TMS, Warrendale, USA, 2000, pp. 23–36.
44. U. Fritsching, V. Uhlenwinkel, K. Bauckhage, U. Urlau, *Chem. Ing. Tech.* **62** (1990) 2, 146 – 147.
45. U. Heck, Dissertation Universität Bremen 1998.

46. I. E. Anderson, R. L. Terpstra, *Mater. Sci. Eng.* **A326** (2002), 101–109.
47. U. Heck, U. Fritsching, K. Bauckhage, *Atomization & Sprays* **10** (2000), 25–46.
48. V. Uhlenwinkel, U. Fritsching, K. Bauckhage, U. Urlau, *Chem. Ing. Tech.* **62** (1990) 3, 228 – 229.
49. U. Fritsching, K. Bauckhage, *PHOENICS J. Comp. Fluid Dynamics* **5** (1992) 1, 81 – 98.
50. A. M. Sizov: *Dipersion of Melts by Supersonic Gas Jets*, Metallurgija Publ. Moskau 1991.
51. U. Fritsching, V. Uhlenwinkel, K. Bauckhage, K. (eds.): *Mater. Sci. Eng.* **A326** (2002) 1.
52. H. Lubanska, *Journal of Metals* **2** (1970) 45 – 49.
53. V. G. Levich, *Physicochemical Hydrodynamics*, Prentice-Hall, New-Jersey 1962.
54. D. Bradley, *J. Phys. D: Appl. Phys.* **6** (1973) 1724–1736;
J. Phys. D: Appl. Phys. **6** (1973) 2267–2272.
55. S. Markus, U. Fritsching, K. Bauckhage, *Mater. Sci. Eng.* **A326** (2002), 122–133.
56. N. Dombrowski, W. R. Johns, *Chem. Eng. Sci.* **18** (1993), 203–214.
57. E. Klar, J. W. Fesko, *Powder Metallurgy Metals Handbook*, 9th ed., vol. 7; American Society for Metals, Materials Park, Ohio, 1984.
58. S. A. Kreczkowski, *Int. J. Multiphase Flow* **6** (1980), 227–235.
59. M. Pilch, M., C. A. Erdmann, *Int. J. Multiphase Flow* **13** (1987) 741–757.
60. W. Samenfink, A. Elsäßer, K. Dullenkopf, Proceedings ICLASS-94, 1994, pp. 156–163.
61. L. P. Hsiang, G. M. Faeth, *Int. J. Multiphase Flow* **18** (1992), 635–652.
62. R. Schmehl: 8. Int. Conference on Liquid Atomization and Spray Systems ICLASS, Pasadena, CA, USA, 2000.
63. P. J. O'Rourke, A. A. Amsden, *The TAB Method For Numerical Calculation of Spray Droplet Breakup*, Los Alamos National Laboratory LA-UR-87-2105 (SAE Technical Paper 872089).
64. R. I. Nigmatulin, *Dynamics of Multiphase Media*, English Edition Editor, **vol. 1, 2**, 1990.
65. Z. Dai, L. P. Hsiang, G. M. Faeth, *Int. J. Multiphase Flow* **18** (1992), 635–652.
66. W. H. Chou, G. M. Faeth, *Int. J. Multiphase Flow* **24** (1998), 889–912.
67. Z. Dai, G. M. Faeth, *Int. J. Multiphase Flow* **27** (2001), 217–236.
68. G. M. Faeth, Proc. ILASS- Europe 2002, Saragossa, Spain, 2002.
69. S. Conelly, J. S. Coombs, J. O. Medwell, *Metal Powder Report* **41** (1986) 9.
70. H. Liu: *Science and Engineering of Droplets: Fundamentals and Applications*, William Andrew Publ., Norwich, USA, 2000.
71. J. Domnick, J. Raimann, G. Wolf, A. Berlemont, M.-S. Cabot, *Proc. ICLASS-'97*, Seoul, Korea 1997.
72. K. Bauckhage, Proc. PM^{2}TEC'98, Las Vegas 1998.
73. M. Krauss, D. Bergmann, U. Fritsching, K. Bauckhage, *Mater. Sci. Eng.* **A 326** (2002), 154–164.
74. J. Ziesenis, K. Bauckhage: Proc. ICLASS 2003, Sorrento Italy, 2003.
75. D. Bergmann, U. Fritsching, C. T. Crowe, Proc. 2nd Int. Conf. Multiphase Flow, Kyoto, Japan, 1995, Vol. 1, SP1 – SP8.
76. S. J. Pien, J. Luo, F. W. Baker, M. K. Chyu: *Koll. SFB 372*, vol. **1**, Universität Bremen 1996, pp. 161 – 168.
77. D. Bergmann, U. Fritsching, K. Bauckhage, *Comp. Fluid Dynamics-J.* **9** (2001), 203–211.
78. H. Liu, Dissertation Universität Bremen 1990.
79. U. Fritsching, H. Liu, K. Bauckhage, *Proc. ICLASS-´91*, Gaithersburg, M.D., USA, NIST SP813 (1991) 491 – 498.
80. M. Krauss, D. Bergmann, U. Fritsching, *Mater. Sci. Eng.* **A326** (2002), 154–164.
81. B. Ruck: *Lasermethoden in der Strömungsmeßtechnik*, AT-Fachverlag, Stuttgart, 1990.
82. E. D. Hirleman: *Part. Part. Syst. Charact.* **13**, 1996, 59–67.
83. H.H. Flögel: Diss., Universität Bremen, VDI Verlag, 1987.
84. J. Tillwick, V. Uhlenwinkel, K. Bauckhage: Proc. ILASS 98, Manchester, 1998, pp. 388–393.
85. J. Tillwick, J. Ziesenis, V. Uhlenwinkel: PM^{2}TEC, Vancouver, 20.-24.6.1999.
86. J. Tillwick: Dissertation, Universität Bremen, 1999.
87. J. Ziesenis, J. Tillwick, V. Uhlenwinkel, K. Bauckhage: Proc. SDMA 2000, Bremen 2000.
88. K. Bauckhage: Proc. PM^{2}TEC'98, Las Vegas, 1998.
89. V. Uhlenwinkel, M. Buchholz, K. Bauckhage: TMS-Meeting, EPD Congress, San Diego, 1999.
90. V. Uhlenwinkel, U. Fritsching, K. Bauckhage, *Proc. ICLASS-91*, Gaithersburg, M.D., USA, NIST SP813 (1991) pp.483 – 490.
91. V. Uhlenwinkel, Dissertation Universität Bremen 1992.

92. C. Kramer, Dissertation Universität Bremen 1997.
93. K. Bauckhage, D. Bergmann, J. Tillwick: *Koll. SFB 372*, **vol. 4**, Bremen 1999, pp. 139–170.
94. M. Buchholz, V. Uhlenwinkel, A. v. Freyberg, K. Bauckhage, *Mater. Sci. Eng.* **A326** (2002), 165–175.
95. G. Trapaga, J. Szekely, *Metall. Trans. B* **22** (1991) 901 – 910.
96. M. Berg, Dissertation, Universität Bremen 1999.
97. M. Bussmann, J. Mostaghimi, S. Chandra, *Phys. Fluids* **11** (1999), 1406–1417.
98. M. Bussmann, S. D. Aziz, S. Chandra, *J. Heat Transfer* **122** (2000), 422.
99. H. Liu, E. J. Lavernia, R. H. Rangel, *Atomization and Sprays* **4** (1994) 369 – 384.
100. J. P. Delplanque, E. J. Lavernia, R. H. Rangel, *Numerical Heat Transfer, Part A* **30** (1996) 1 – 18.
101. U. Fritsching, H. Zhang, K. Bauckhage, *Proc. Int. Conf. Spray Forming ICSF-2*, Swansea UK 1993.
102. M. Gutierez-Miravete, E. J. Lavernia, G. M. Trapaga, J. Szekely, *Int. J. Rapid Solidification* **4** (1988) 125 – 150.
103. O. Meyer, U. Fritsching, K. Bauckhage, *Int. J. Thermal Sci.* **42** (2003), 153–168.
104. O. Meyer, A. Schneider, V. Uhlenwinkel, U. Fritsching, *Int. J. Thermal Sci.* **42** (2003), 561–569.
105. H. Zhang, Dissertation Universität Bremen 1994.
106. U. Fritsching, H. Zhang, K. Bauckhage, *steel research* **65** (1994) 7, 273 – 278.
107. U. Fritsching, D. Bergmann, K. Bauckhage, *Proc. ICLASS-97*, Seoul, Korea 1997.
108. Z. Djuric, P. Newberry, P. S. Grant, *Modelling Simul. Mater. Sci. Eng.* **7** (1999) 553–571.
109. Z. Djuric, P. S. Grant, *Modelling Simul. Mater. Sci. Eng.* **9** (2001) 111–127.
110. W. D. Cai, E. J. Lavernia, *Mater. Sci. Eng. A* **226** - 228 (1997) 8–12.
111. A. Lawley: *Atomization - The Production of Metal Powders*, Metal Powder Industries Federation, Princeton, USA, 1992.
112. M. Buchholz, Dissertation Universität Bremen 2002.
113. V. Uhlenwinkel, N. Ellendt, M. Walter, J. Tockner: Int. Conference on Continuous Casting of Non-ferrous Metals, Neu Ulm, Germany, 2006.
114. J.H. Hattel, PhD Dissertation, Technical University Denmark, Lyngby 1999.
115. J. H. Hattel, N. H. Pryds, T. B. Pedersen, A. S. Pedersen: Proc. SDMA 2000, Bremen, pp. 803–812.
116. D. Bergmann, Dissertation Universität Bremen 2000.
117. R. S. Minisandram, R. M. Forbes Jones, K. M. Kelkar, S. V. Patankar, W. T. Carter Jr., *Mater. Sci. Eng.* **A326** (2002), 184–193.
118. T.B. Pedersen, Ph. D. thesis Technical University of Denmark, 2003.
119. K. Petersen, A. S. Pedersen, N. Pryds, K. A. Thorsen, J. L. List, *Mater. Sci. Eng.* **A326** (2002), 40–50.
120. N. Pryds, J. H. Hattel, T. B. Pedersen, *J. Thorborg, Acta Mater.* **50** (2002), 4075–4091.
121. D. Bergmann, U. Fritsching, K. Bauckhage K: *Int. J. Thermal Sci.* **39** (2000), 53–62.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Aerogels

Standard Article

Nicola Hüsing[1] and Ulrich Schubert[2]

[1]Anorganische Chemie I, Universität Ulm, 89081 Ulm, Germany

[2]Institut für Materialchemie, Technische Universität Wien, A-1060 Wien, Austria

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.c01_c01.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Synthesis
2.1. Network Formation and Structure
2.1.1. Inorganic Aerogels
2.1.1.1. SiO_2 Aerogels
2.1.1.2. Metal Oxide Aerogels
2.1.2. Inorganic/Organic Hybrid Aerogels
2.1.3. Organic Aerogels
2.2. Drying Methods
2.2.1. Supercritical Drying
2.2.1.1. Drying in Organic Solvents
2.2.1.2. Supercritical Drying with CO_2
2.2.2. Freeze Drying
2.2.3. Ambient Pressure Drying
2.3. Modification of Aerogels after Drying
3. Properties
3.1. Structural Properties
3.2. Physical and Chemical Properties
3.2.1. Optical Properties
3.2.2. Mechanical and Acoustic Properties
3.2.3. Thermal Conductivity
3.2.4. Hydrophobicity
4. Applications
4.1. Çerenkov Detectors
4.2. Thermal Insulation
4.3. Catalysis
4.4. Application as Storage Media
4.5. Electrical and Electronic Applications
4.6. Acoustic and Mechanical Applications
4.7. Optical Applications
4.8. Other Applications
5. Future Perspectives

Aerogels are unique among solid materials. They have extremely low densities (up to 95 % of their volume is air), large open pores, and a high inner surface area. This combination of properties results in interesting physical properties, e.g., for silica aerogels extremely low thermal conductivities and low sound velocities are combined with optical transparency. Aerogels are typically prepared via wet chemical processes, such as sol – gel processing, and thus their pore system is

initially filled with liquid. Only special drying techniques allow for exchange of the pore liquid against air while maintaining the filigrane solid framework. The most common drying technique is supercritical extraction, and sometimes chemical modification of the inner surface can be used to facilitate drying. The structure of the gel network, and thus the physical properties of aerogels, decisively depends on the choice of the precursors and the chemical reaction parameters for preparing the gels. Therefore, the later materials properties can be deliberately designed in the starting reaction solution.

[Top of Page]

1. Introduction [1-11], [12-14]

Aerogels are highly porous solid materials with extremely low densities (bulk densities 0.004 – 0.500 g/cm^3), large, open pores, and high specific surface areas. This results in interesting physical properties, such as extremely low thermal conductivity and low sound velocity, combined with high optical transparency. Aerogels have been differently defined:

1. All materials prepared from wet gels by supercritical drying were called aerogels, irrespective of their structural properties. With the development of new drying techniques this definition is no longer appropriate.
2. Materials in which the typical pore structure and network are largely maintained when the pore liquid of a gel is replaced by air are called aerogels (this definition is used in this article). However, it is not always clear to what extent the structure is maintained.

Materials are porous if they contain cavities, channels, or interstices that are deeper than they are wide. The pores may be regularly arranged, as in molecular sieves such as zeolites (⟶ Zeolites). However, the more common situation is an irregular pore structure. The physical properties of a porous solid and its reactivity are effectively influenced by the type, shape, and size of the pores. Some technical terms are defined in Table 1.

The unique materials properties of aerogels result from the special arrangement of their solid network, which is schematically shown in Figure 1 for an SiO_2 aerogel [16]. The structures of aerogels are characterized by readily accessible, cylindrical, branched mesopores. Aerogels can be obtained as monoliths, granulates, or powders.

Table 1. Definition of terms for porous solids

Term	Explanation
Porosity of a solid	ratio of the total pore volume to the volume of the particles or powder
Pore shape	ink bottle, cylindrical, funnel, slit-shaped
Accessibility of the pores	closed, blind, through pores
Pore size and dominating transport mechanisms [15]	*micropores:* < 2 nm, activated transport
	mesopores: 2 – 50 nm, Knudsen diffusion, surface diffusion along the pore walls, capillary transport
	macropores: > 50 nm, molecular diffusion
Density	skeletal density: density of the solid network
	bulk density: mass per total volume (= solid phase + closed pores + open pores)

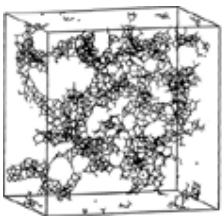

Figure 1. Schematic structure of a silica aerogel (reprinted from [16] with permission of Elsevier Science Publishers).

[Top of Page]

2. Synthesis

The network and hence the pore structure of gels is formed by the condensation of small (polymeric or colloidal) primary particles with a diameter of 1 – 3 nm. Their generation and aggregation is controlled by chemical processes, usually the sol – gel process [17-24]. In principle, the structures of all covalently bonded gels lie between those of colloidal gels (colloidal primary particles) and polymeric gels (polymeric primary particles), which constitute the two extremes with regard to both the microstructure and the resulting properties (Figure 2).

Figure 2. Two-dimensional structures of aerogels a) Colloidal; b) Polymeric; c) With a rough surface (see Section SiO_2 Aerogels) (Reprinted from [50] with permission of the Materials Research Society)

In a *sol*, particles with diameters in the range of 1 – 1000 nm are dispersed in a liquid. A *gel* consists of a sponge-like, three-dimensional solid network whose pores are filled with another substance, usually a liquid. When gels are prepared by hydrolysis and condensation of metal or metalloid alkoxides or other hydrolyzable metal compounds (via the sol stage), the pore liquid mainly consists of water and/or alcohol. The resulting “wet” gels are therefore called aquagels, hydrogels, or alcogels. *Aerogels* are obtained when the pore liquid is replaced by air without decisively altering the network structure or the volume of the gel body. *Cryogels* are obtained when the pore liquid is removed by freeze drying. A *xerogel* is formed by conventional drying of the wet gels, that is, by increasing the temperature or decreasing the pressure with concomitant large shrinkage (and mostly destruction) of the initially uniform gel body (Figure 3). The shrinkage of a gel body upon evaporation of the pore liquid is caused by capillary forces acting on the pore walls as the liquid retreats into the gel body. This results in the collapse of the filigree, highly porous inorganic networks of the aquagels or alcogels. Therefore, other drying methods must be used to prepare aerogels (Section Drying Methods).

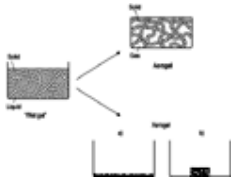

Figure 3. Shrinkage upon drying of a wet gel body to give an aerogel (top) and to give a xerogel (bottom) as a powder (a) or as monolith (b)

2.1. Network Formation and Structure

Highly porous three-dimensional gel networks can be obtained for inorganic, inorganic/organic, and purely organic systems under controlled conditions although the solids content is typically only 1 – 15 vol %.

In colloidal gels, dense colloidal particles are interconnected like a string of pearls (Figure 2a). In polymer gels, linear or branched polymer chains are formed by condensation of small clusters (Figure 2b). In the sol – gel literature and in this article, the term cluster does not mean metal clusters, as is usual in chemistry, but instead the primary oligomeric species formed during sol – gel processing. Their structure can be chainlike, cyclic, or three-dimensional. There are several theoretical models for the aggregation of the clusters to three-dimensional networks [17], [25]. Each aerogel has its own structural characteristics because the microstructure strongly depends on the preparation conditions.

Aerogel networks are usually formed from colloidal particles. There are two principal possibilities for the formation of gels:

- Solid gels are dissolved (peptized) and the sols thus obtained (colloids) are reaggregated under different conditions.
- Sol particles are formed by chemical reactions from molecular precursors. Most aerogel syntheses follow this route. For example, the reaction of metal or metalloid alkoxide groups (M–OR) with water results in the formation of hydroxyl groups (M–OH), which can then condense to form M–O–M units.

The production of inorganic and inorganic/organic aerogels is shown schematically in Figure 4. Some parameters by means of which the sol – gel process and therefore the properties of the resulting aerogels can be influenced and varied are also listed. Organic aerogels are produced similarly.

Figure 4. General scheme for the preparation of aerogels by sol – gel processing and some typical parameters for variation

The solvent plays an important role in sol – gel processes. It not only serves to homogenize the precursors in the initial stage, but also strongly influences the particle- and network-forming reactions due to its polarity and viscosity. In the preparation of aerogels it has an additional function, as shown in Figure 4. Since gelation results in only a marginal change in volume, and the drying process during aerogel production is performed (by definition) such that shrinkage is minimal, the volume of the aerogel body (and hence its density) is determined by the volume of the reaction solution. Therefore, the density of aerogels is simply modified by varying the precursor concentration in the starting solution.

The sol – gel transition (gel point) is reached when a continuous network is formed. However, the chemical reactions are not finished at this point. First, the pore liquid initially is a sol, that is, it contains condensable particles or even monomers,

which slowly condense onto the existing network. Second, neighboring M–OH or M–OR groups can undergo condensation reactions (aging), because the gel network is still very flexible. Furthermore, hydrolysis and condensation reactions are in principle reversible. Therefore, mass is dissolved from thermodynamically unfavorable regions. The solutes condense in thermodynamically more favorable regions, especially in pores, crevices, particle necks, etc. (Ostwald ripening). This process results in the reduction of the net curvature, the disappearance of small particles, and the filling of small pores. Aging and ripening increase the stiffness of the gels (Fig. 4). Controlled aging is therefore an important aspect for the reproducible preparation of aerogels.

2.1.1. *Inorganic Aerogels*

2.1.1.1. *SiO_2 Aerogels*

The term SiO_2 gel (analogously for other gels) is only used to characterize the type of inorganic skeleton. The SiO_2 gels and the corresponding aerogels often have the composition $SiO_x(OH)_y$ or, if they are prepared from alkoxides, $SiO_x(OH)_y(OR)_z$, where the values of *y* and *z* can be (but need not be) rather high.

Most laboratory preparations of silica aerogels use tetraalkoxysilanes $Si(OR)_4$ as the silica source. The pore liquid of the gels formed after addition of a defined amount of water (Eq. 1) then mainly consists of the alcohol used as solvent and that produced by the hydrolysis reaction.

$$Si(OR)_4 + H_2O/ROH \rightarrow [SiO_2 \cdot H_2O/ROH]$$

The chemical reactions during sol – gel processing of alkoxysilanes can be formally described by three equations (Eq. 2). Hydrolysis reactions take place in parallel with condensation reactions during all steps of the sol – gel process. Therefore, all intermediate species still contain Si-OR and/or Si-OH groups.

$$\begin{array}{lcll} \equiv Si{-}OR + H_2O & \xrightleftharpoons{\text{acid or base}} & \equiv Si{-}OH + ROH & \text{Hydrolysis} \\ \equiv Si{-}OH + HO{-}Si\equiv & \rightleftharpoons & \equiv Si{-}O{-}Si\equiv + H_2O & \\ \equiv Si{-}OR + HO{-}Si\equiv & \rightleftharpoons & \equiv Si{-}O{-}Si\equiv + ROH & \text{Condensation} \end{array} \quad (2)$$

Parameters which influence the hydrolysis and condensation reactions and whose deliberate variation are used for “materials design” are shown in Figure 4. A variety of physicochemical measurements and some computer simulations were performed to determine the influence of these parameters on network formation of SiO_2 aerogels [7-11], [25-31]. Kinetic models for the growth of gel networks were postulated on the basis of small-angle X-ray scattering (SAXS) data [32], [33]. Two of the most important models are shown in Figure 5.

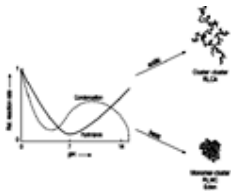

Figure 5. Dependence of the hydrolysis and condensation reaction on pH, and derived kinetic growth models of gel structures. RLCA = reaction limited cluster aggregation, RLMC = reaction limited monomer cluster growth (after [32], [25])

The kind of inorganic network formed by the condensation reactions not only depends on the absolute rates of the individual reactions but also on the relative rate of hydrolysis and condensation reactions [34], [35]. Acidic conditions (pH 2 – 5) favor hydrolysis, and condensation is the rate-determining step. A great number of monomers or small oligomers with reactive Si–OH groups are simultaneously formed. Under these conditions, reactions at terminal silicon atoms are favored, and chains with few branches are formed. This results in polymerlike networks with small pores. This process is called reaction limited cluster aggregation (RLCA).

Hydrolysis is the rate-determining step under basic conditions, which favor reaction at central silicon atoms of an oligomer. The resulting network has a particulate character with large particles and large pores (“colloidal” gels). Hydrolyzed species are immediately consumed because of the faster condensation reaction. Condensation of clusters with each other is relatively unfavorable [17]. Therefore, the clusters mainly grow by the condensation of monomers. This process is called reaction limited monomer cluster growth (RLMC) or Eden growth [36]. It would ideally result in uniformly dense structures. However, for tetraalkoxysilanes this simple model must be somewhat modified. Not all the four branching points are equal because only partially hydrolyzed $Si(OR)_{4-x}(OH)_x$ is initially formed from $Si(OR)_4$. Therefore, unchanged Si–OR groups are present at the growth front, which leads to inhomogeneous condensation, and rough structures are obtained (“poisoned” Eden growth; Fig. 2c).

Preparation of aerogels from polymeric gels is more difficult, because diffusion processes are strongly inhibited by the smaller pores. Complete removal of the pore liquid is therefore more difficult and results in a larger shrinkage during drying [37]. For this reason, silica aerogels are usually prepared by base-catalyzed reaction of tetramethoxysilane (TMOS) or tetraethoxysilane (TEOS), mostly with ammonia as the catalyst.

A modification of this procedure is to pre-hydrolyze $Si(OR)_4$ with a small amount of water under acidic conditions. This

results in the formation of small silicic acid clusters. In a second step, a defined amount of aqueous acid or base is added [38]. Networks formed by this two-step procedure have a structure similar to the RLCA model, that is, they have a polymeric character. The probable explanation is that the reactive clusters formed in the first step as structure forming units are responsible for the formation of the network, regardless of the kind of catalyst in the second step. The main difference caused by the kind of catalyst in the second step is the stiffness of the resulting network. Base catalysis in the second step results in stiffening, which stabilizes the gels. With this two-step procedure a more deliberate control of the microstructure and hence the particle and pore sizes of SiO_2 gels is possible. For example, while the density of aerogels made by one-step base- or acid-catalyzed reactions is restricted to the range of 0.030 – 0.300 g/cm^3, the production of an aerogel with a density of only 0.004 g/cm^3 was achieved by the two-step process [39], [40].

In a modification of the two-step procedure, oligomeric polyethoxysiloxanes of a certain size are prepared first from TEOS by addition of a defined amount of water. These defined oligomers were then used as precursors for sol – gel processing to form the three-dimensional network [41], [42].

For technical applications in which large amounts of aerogels are needed, alkoxysilanes are too expensive. Therefore, *water glass*, an aqueous sodium silicate solution, is used as the silica source. The sodium silicate solution is ion-exchanged, and the resulting silicic acid solution is gelled by changing the pH to acidic conditions [43], [44].

Silicic acid can also be transferred into an organic solvent in which condensation is then performed under acidic, neutral, or basic conditions. The resulting wet gels are also potentially suitable for the production of aerogels [45].

For the commercial production of Basogel, a two-step procedure was applied: The sodium silicate solution is mixed with sulfuric acid in the first step. Small hydrogel or aquagel droplets are formed by spraying. The metal salts are then extracted, and the water is replaced by an organic solvent [46]. The gel droplets are then supercritically dried.

The macroscopic properties of SiO_2 aerogels differ over a wide range, because of the differences in the structure of the primary particles and network formation. The structural data of SiO_2 aerogels listed in Table 2 therefore only serve as examples.

Table 2. Properties of SiO_2 aerogels

Property	Range	Typical value
Bulk density, g/cm^3	0.003 – 0.500	0.100
Skeletal density, g/cm^3	1.700 – 2.100	
Porosity, %	80 – 99.8	
Mean pore diameter, nm	20 – 150	
Specific surface area, m^2/g	100 – 1600 [47]	600
Refractive index	1.007 – 1.24	1.02
Thermal conductivity (in air, 300 K), W m^{-1} K^{-1}	0.017 – 0.021	0.020
Modulus of elasticity *E*, MPa	0.002 – 100	1
Sound velocity c_L, m/s	< 20 – 800	100
Acoustic impedance *Z*, kg m^{-2} s^{-1}		10^4

2.1.1.2. *Metal Oxide Aerogels*

The principles for network formation of nonsilicate inorganic gels are the same as for SiO_2 gels. Aqueous salt solutions or molecular precursors in organic solvents, again mostly alkoxides, can be employed for sol – gel processing.

Metal alkoxides are much more reactive towards water than alkoxysilanes. This is due to the lower electronegativity and higher Lewis acidity as well as the possibility of increasing the coordination number [48]. For example, the reactivity of tetravalent alkoxides in hydrolysis reactions decreases in the order $Si(OiPr)_4 \ll Sn(OiPr)_4$, $Ti(OiPr)_4 < Zr(OiPr)_4 < Ce(OiPr)_4$. The reactivity of some metal alkoxides is so high that precipitates are spontaneously formed upon addition of water. Whereas the reactivity of alkoxysilanes has to be promoted by catalysts, the reaction rates of metal alkoxides must be moderated to obtain gels instead of precipitates. The most common method is the addition of acetic acid or acetylacetone to the precursor solution. This results in partial substitution of the alkoxy groups by acetate or acetylacetonate. Other bidentate anionic ligands (L) can also be used [49]. The complexes $[\{M(OR)_y(L)_x\}_n]$ have a different reactivity, structure, and functionality compared to the unsubstituted alkoxide $M(OR)_{y+x}$ depending on the type and number of ligands. Complexation allows the chemical design of the precursors, so that the properties of the sol – gel materials can be deliberately influenced.

Compared to aerogels with a silicate network, only little work has been done to elucidate the network formation and structure of metal oxide gels. One of the most important differences is the possibility of forming crystalline primary particles. Generally, crystallinity is favored by a large excess of water in the hydrolysis reaction.

Alumina. The preferred precursors for the synthesis of Al_2O_3 aerogels are $Al(OsecBu)_3$ (ASB) and $Al(OtBu)_3$ (ATB) [50-54]. ASB has the advantage of being soluble in 2-butanol, while other aluminum alkoxides are insoluble in the corresponding alcohol. There are different possibilities for handling this problem [55]. The alkoxide can be dissolved in benzene followed by addition of pure water (two liquid phases are formed), or water dissolved in the corresponding alcohol. Alternatively, a dispersion of the alkoxide in the corresponding alcohol can be prepared and then water is added.

Titania and Zirconia. For the preparation of TiO_2 and ZrO_2 aerogels, the propoxides and butoxides of the metals are mainly used as precursors [50], [56-60]. As for the alumina gels, the amount of water added for hydrolysis is decisive for the structure of zirconia aerogel networks. Substoichiometric addition of water results in an amorphous network, while with an excess of water the structure is composed of crystalline, monoclinic ZrO_2 particles [3], [50]. When $Zr(OnPr)_4$ was gelled in 1-propanol in the presence of acetic acid, the acetate ligand remained in the aerogel [61].

TiO_2 aerogels can be prepared totally amorphous or with a network of anatase primary particles [62-64]. It was shown by a combination of SAXS, small-angle neutron scattering (SANS), BET measurements, and electron microscopy that mesoparticles of about 50 nm in diameter are formed by aggregation of the crystalline nanoparticles of about 5 nm diameter. The mesoparticles aggregate to give a three-dimensional network. A branched, polymerlike fractal structure was concluded from SAXS data on amorphous and partially crystalline ZrO_2 aerogels consisting of small primary particles with an average diameter of 5.2 nm (partially crystalline) or 2.5 nm (amorphous) [65], [66].

Other single-component oxide aerogels are listed in Table 3.

Table 3. Other inorganic aerogels, precursors for their preparation, and potential applications

Aerogel	Precursor*	Application
V_2O_5	$[VO(OiPr)_3]$ [67], [68]	cathode in lithium batteries
	$[VO(OEt)_3]$ [69-71]	
Cr_2O_3	$[Cr(NO_3)_3]$ [72]	oxidizing properties, fluorination of hydrocarbons with HF
	$CrCl_3$ [73]	
	$Cr(OAc)_3$ [74], [75]	
Fe_2O_3	$FeCl_3$ [72], [73]	
	$[Fe(acac)_3]$ [76-78]	
MoO_2	$[MoO_2(acac)_2]$ [50], [79], [80]	electrocatalysis
Nb_2O_5	$Nb(OEt)_5$ [81-83]	isomerizations, solid acid
MnO_2	$NaMnO_4 \cdot x\ H_2O$ [68], [84]	charge-storage applications
NiO	$NiCl_2 \cdot 6\ H_2O$ [85]	semiconductor, catalytic properties
CdS	$Cd(NO_3)_2 \cdot x\ H_2O$ [86]	photovoltaic and sensing applications

* acac = acetylacetonate.

A newer sol – gel route has opened access to a wide variety of transition and main group metal oxide aerogels such as Cr_2O_3, Fe_2O_3, Al_2O_3, In_2O_3, Ga_2O_3, SnO_2, HfO_2, ZrO_2, Nb_2O_5, Ta_2O_5, and WO_3 by use of epoxides as gelation agents [85], [87-89].

Binary and Ternary Oxide Aerogels. Compared to other methods for the preparation of multicomponent oxides, the sol – gel method offers the best control over the resulting materials properties. This is achieved by variation of the processing parameters [79], [90], [91].

When precursors with similar hydrolysis and condensation rates are mixed, mixed metal oxides are obtained which are homogeneous on a molecular scale. With larger differences in reaction rates, the microstructure of the product becomes more heterogeneous, and phase separation may even occur. In particular, the so-called core/shell phenomenon is observed. The faster reacting compound forms sol particles which are coated by the slower reacting component. There are several possibilities for avoiding this effect: The reaction rate of the faster reacting component can be moderated by using bidentate ligands. Alternatively, the slower reacting precursor can be prehydrolyzed and the faster reacting component added afterwards [92]. The network in the aerogels thus prepared is analogous to that of the one-component oxide aerogels.

Another option is the impregnation of a preformed silica alkogel with a metal alkoxide in a supercritical medium, as shown for titanium alkoxides [93].

Selected examples are listed in Table 4. Mixed oxide aerogels for catalytic applications are summarized in review articles [3], [5].

Table 4. Examples of binary and ternary oxide aerogels

Aerogel	Precursor*	Application
TiO_2 – SiO_2	[Ti(O*i*Pr)$_4$ + acac] + Si(OMe)$_4$ [94], [95]	epoxidation of olefins
Al_2O_3 – SiO_2 [96]	[$AlCl_3$ + ethylene oxide] + Si(OEt)$_4$ [73]	catalyst support
	Al(OAc)$_3$, Al(acac)$_3$	
3 Al_2O_3 – 2 SiO_2 (mullite)	(ASB + β-diketone) + Si(OEt)$_4$ [97]	preparation of high-purity ceramics
Fe_2O_3 – SiO_2 [98]	Fe(acac)$_3$ + Si(OMe)$_4$	Fischer – Tropsch syntheses
Fe_2O_3 – Al_2O_3	Fe(acac)$_3$ + ASB	
V_2O_5 – MgO	V(acac)$_3$ + Mg(OMe)$_2$ [99]	ammonia synthesis
PbO – Al_2O_3	Pb(OAc)$_2$ + ASB [100]	nitroxidation of hydrocarbons to nitriles
PbO – ZrO_2	Pb(OAc)$_2$ + Zr(O*i*Pr)$_4$ [101]	
BaO – Al_2O_3	Ba[Al(O*s*Bu)$_4$]$_2$ + ASB	car exhaust catalysis
	Ba(1,3-butanediolate) + ASB [102]	
x Li_2O – (1 – *x*) B_2O_3	B(OBu)$_3$ + LiOMe [103]	
$PbTiO_3$	Pb(OAc)$_2$ + Ti(O*i*Pr)$_4$ [104], [105]	piezoelectric lead zirconate titanate (PZT) aerogels
$BaTiO_3$, $SrTiO_3$	Ba or Sr and Ti(O*i*Pr)$_4$ [106]	
Mn-FeO_x	$FeCl_3 \cdot 6\ H_2O$ + $MnCl_2 \cdot 4\ H_2O$ [107], [108]	superparamagnetic material
NiO – Al_2O_3 – MgO	Ni(OAc)$_2$ + ASB + Mg(OMe)$_2$	nitroxidation
NiO – SiO_2 – MgO	Ni(OAc)$_2$ + Si(OR)$_4$ + Mg(OMe)$_2$ [109]	
2 MgO – 2 Al_2O_3 – 5 SiO_2 (cordierite-like)	Si(OEt)$_4$ + ASB + Mg(NO$_3$)$_2$ [61]	

* acac = acetylacetonate.

SAXS investigations of Al_2O_3 – SiO_2 (mullite) aerogels, prepared from the alkoxides in the presence of bidentate ligands again lead to the postulation of a RLCA-related network [97]. Si/O building blocks are partially incorporated into the alumina clusters and partially integrated onto the surface of these particles. The SiOH groups at the surface of the clusters serve as condensation sites between the particles.

X-ray diffraction (XRD) showed that SiO_2 aerogels are always amorphous, while TiO_2, Al_2O_3, and ZrO_2 can also be partially crystalline [50]. In mixed oxide aerogels, the SiO_2 and Al_2O_3 portions are always amorphous [77], [78], [109-111]. However, crystalline aluminates are observed (for example, $NiAl_2O_4$). Selected structural data of nonsilicate aerogels are listed in Table 5. They vary decisively with the preparation conditions.

Table 5. Typical structural data of some nonsilicate aerogels

Aerogel	Density, g/cm^3	Porosity, %	Pore radius, nm	BET surface area, m^2/g	Particle morphology, nm
TiO_2 [112]	0.3 – 1	78 – 90	1 – 25	316 – 690	
ZrO_2 [50], [61], [69-71], [83]	0.2 – 0.3	84 – 96	10	81 – 480	2.5 – 5.2
Al_2O_3 [50], [54], [113]	0.13 – 0.18		5	123 – 616	22 – 25, platelike
Al_2O_3 – SiO_2 [97], [114], [115]	0.06 – 0.21		12		1 – 5 *
V_2O_5 [69], [116]	0.04 – 0.1	96		140 – 400	< 10 nm, fibrous

$V_2O_5 – GeO_2$ [70]	0.08	96			
Cr_2O_3 [110]	0.15 – 0.54			516 – 785	
$PbO – TiO_2$ [104]	0.7	77	1 – 4	260	35 – 55

* Controllable by means of the amount of added β-diketone [115].

Metal in Metal Oxide Aerogels. A second metal component in a gel need not necessarily result in a binary oxide aerogel. The use of alcohols for supercritical drying can result in the conversion of easily reducible oxides into metals, because alcohols have reducing properties. Metal particles can also be generated by subsequent reduction of the aerogels with hydrogen at high temperature or by introducing hydrogen into the autoclave during supercritical drying [117].

Aerogels are particularly well suited as carriers for catalytically active metals because of their high porosity. Detailed discussions can be found in review articles on aerogels as catalysts [3], [5].

For the incorporation of metal components into the gel matrix, previously formed aerogels can be impregnated with solutions of the corresponding metal salts (two-step process) [118], [119]. Alternatively, metal salts can be added to the precursor solution, which results in the incorporation of the metal compounds in the gel during sol – gel processing [72], [120]. Table 6 gives some examples of metal-doped aerogels and their application in catalytic reactions.

Table 6. Examples of metal-doped aerogels

Aerogel	Precursor for the metal	Application
1. Impregnation		
$Cu – SiO_2$, $Cu – MgO$	$[Cu(NH_3)_4(OH)_2]$ in alcohol [118]	
$Cu – ZnO – Al_2O_3$	$Cu(OAc)_2$ [121]	hydrogenation
$Pt – MoO_2$	$H_2[PtCl_6]$ in MeOH [119]	hydrogenation
$Pd – Al_2O_3$	$PdCl_2$ [122], [123]	car exhaust catalysis
$Pd – CeO_2$,		
$Pd – BaO – Al_2O_3$		
2. Sol – Gel method		
$Pt – TiO_2$	$PtCl_4$, $[Pt(acac)_2]$, $(NH_4)_2[PtCl_6]$ [124]	hydrogenation
$Pd – Al_2O_3$	$[Pd(OAc)_2]$ [125]	car exhaust catalysis
$V – SiO_2$	$V(OiPr)_3$	
$Cu – SiO_2$	$Cu(OAc)_2$ [126]	
$Pt – Al_2O_3$	$H_2[PtCl_6]$ [127]	dehydrocyclisation

2.1.2. *Inorganic/Organic Hybrid Aerogels*

The goal of modifying oxide aerogels with organic groups is to supplement new properties without influencing the existing positive properties, such as good thermal insulation, transparency, and high surface area. For example, the hydrophobicity and elastic properties of SiO_2 aerogels can be improved relative to unmodified SiO_2 aerogels by incorporation of organic groups [128-130], or new applications can be envisioned by the integration of functional organic groups, for example, in catalysts or sensors.

There are only limited possibilities for post-synthesis doping or modification of aerogels with organic compounds (Section Modification of Aerogels after Drying). A more general route is the integration of organic entities in sol – gel processing.

Embedding molecules in gels without chemical bonding (Fig. 6 b) is achieved by dissolving these molecules in the precursor solution. The gel matrix is formed around them and traps them. The doped wet gels can, in principle, be converted to aerogels. However, the probability is very high that the organic groups are leached during the generally applied supercritical drying and the rinsing processes connected with it. For example, a fluorescent dye was incorporated in the gel network, but the portion retained during aging and drying was relatively low [131]. C_{60}/C_{70} was also incorporated into SiO_2 aerogels by this method [132]. Larger entities such as biomolecules (e.g., lipase) and microorganisms (e.g., *E. coli*) can be immobilized on silica aerogels and have great potential as biosensors [133], [134].

Aerogels with interpenetrating, but not interconnected, inorganic and organic networks (Fig. 6 a) were produced by generating the organic polymer in situ during the sol – gel reaction by radical polymerization of a vinyl monomer followed by supercritical drying with CO_2 of the dual-network gel. Attempts to perform the sol – gel process in solutions of organic polymers mostly failed, because the polymer was leached during supercritical drying. There was only success with organic

polymers capable of forming hydrogen bonds to the surface silanol groups of the inorganic matrix, such as poly-2-vinylpyridine (PVP) [135].

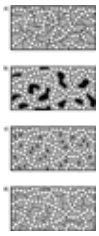

Figure 6. Inorganic/organic hybrid networks from sol – gel processing a) Interpenetrating inorganic and organic networks; b) Incorporation of organic molecules (enzymes, dyes, etc.); c) Organic modification (functionalization) of oxidic materials; d) Dual networks

The more important organic modifications of oxide aerogels are based on covalent bonding of the organic groups (Figure 6 c, d). In silicate systems, compounds of the type $R'Si(OR)_3$ are usually employed, where R′ is the organic group that modifies the inorganic gel. In nonsilicate systems, the organic groups can be introduced by means of bidentate ligands (Section Metal Oxide Aerogels). The preparation of organically modified xerogels is described in review articles [48], [49], [136].

Silica aerogels modified by nonfunctional or functional organic groups are prepared by sol – gel processing of $R'Si(OR)_3/Si(OR)_4$ mixtures (Eq. 3) followed by supercritical drying of the wet gels.

$$xR'Si(OR)_3 + (1-x)Si(OR)_4 + H_2O/ROH \rightarrow R'_xSiO_{(2-0.5x)} \cdot H_2O/ROH$$

When the organic groups R′ have no or only weakly basic properties, such as unsubstituted alkyl or aryl groups, or chloro, thio, phosphino, carbamato, glycidoxy or methacryloxyalkyl groups, sol – gel processing of the $R'Si(OR)_3/Si(OR)_4$ mixtures under base-catalyzed conditions is a two-step process, similar to the formation of core/shell multicomponent oxide aerogels (Section Metal Oxide Aerogels). The basic gel network is almost exclusively built by hydrolysis and condensation of $Si(OR)_4$, because reaction of $Si(OR)_4$ is faster than that of $R'Si(OR)_3$. Only in the second stage of the process are the $R'SiO_{3/2}$ units condensed onto the inner surface of the existing silica gel network. Inorganic/organic hybrid aerogels prepared from $R'Si(OR)_3/Si(OMe)_4$ mixtures therefore have the same structural features as those prepared from $Si(OR)_4$ alone, and the typical aerogel properties are retained. As long as the group R′ has no basic properties, the nature of R′ has no major influence on network formation and network structure, and only the fractal dimension of the primary particles is influenced by the polarity of the functional group [137-141]. Increasing the $R'Si(OR)_3$ fraction leads to larger primary particles and therefore lower specific surface areas. These structural phenomena result from the larger amount of water and catalyst that act on $Si(OR)_4$ in the first stage of the reaction due to the delayed reaction of $R'Si(OR)_3$.

When R′ in the $R'Si(OR)_3/Si(OR)_4$ mixture has basic properties, as in $(RO)_3Si(CH_2)_3NR'_2$ ($NR'_2 = NH_2$ or $NHCH_2CH_2NH_2$), both silanes are involved in the build-up of the gel network, and the two-stage process is no longer observed. The basic groups catalyze the sol – gel reactions [142], [143].

In another approach, the morphology (pore size, surface area) of the aerogels is tailored by linking $Si(OR)_3$ groups with an organic spacer. Compounds of the type $(RO)_3Si–X–Si(OR)_3$ with variable length and composition of the spacer X (e.g., saturated or unsaturated hydrocarbon chains, polyaryls) have been used. An inorganic oxidic network is formed during sol – gel processing in which the organic spacers X bridge the inorganic building blocks. When precursors of the type $(RO)_3Si(CH_2)_nSi(OR)_3$ (n = 2, 6, 8, 10, 14) were processed without the addition of $Si(OR)_4$ both under basic and acidic conditions, the length of the spacer n had a decisive influence on the resulting aerogel structure [144]. Ethylene- and hexamethylene-modified gels were highly porous and had high specific surface areas (about 700 m^2/g). The specific surface area decreased with increasing length of the spacer.

2.1.3. *Organic Aerogels*

Purely organic aerogels were synthesized by polymerization of multifunctional organic monomers in dilute solution, followed by supercritical drying [145-151]. Gelation of resorcinol/formaldehyde (RF) and melamine/formaldehyde (MF) mixtures was initiated by Na_2CO_3 or NaOH. The gels were then treated with dilute acid to increase the degree of cross-linking and thus stabilize the network. The most important parameter is the catalyst concentration or the pH of the solution. The structure and properties of the organic gels, such as density, specific surface area, and particle and pore sizes, can be deliberately influenced by these variables.

Resorcinol is a trifunctional monomer which can add up to three equivalents of formaldehyde. These substituted resorcinol derivatives then condense by formation of nanosized clusters, which then cross-link via the surface groups (for example, $–CH_2OH$). For the polymerization of melamine with formaldehyde, either the monomers or melamine – formaldehyde polymers of low molecular mass can be used [152].

Newer developments in the area of organic aerogels are based on the polymerization of furfural with a phenolic Novolak resin (prepared by reaction of phenol with formaldehyde) to give the so-called PF gels and polymerization based on polyaddition reactions of polyols and polyisocyanates to give polyurethane gels. A decisive advantage compared with RF or MF polymerization is that alcoholic solutions can be used. The laborious solvent exchange before drying is thus avoided

[153], [154].

Despite the structural differences originating from the different precursors, organic aerogels have some common properties, such as small, accessible pores (≤ 50 nm), high specific surface areas (400 – 1000 m^2/g), and a solid matrix which is composed of interconnected colloidal particles or polymeric chains of about 10 nm in diameter. Table 7 lists some representative data for organic aerogels.

Table 7. Some typical structural data and properties of organic aerogels

	Resorcinol – formaldehyde (RF)	Melamine – formaldehyde (MF)	Phenol – furfural (PF) [153]
Density, g/cm^3	0.03 – 0.6	0.10 – 0.80	0.10 – 0.25
Surface area, m^2/g	350 – 900	875 – 1025	385
Pore diameter, nm	50	50	
Particle morphology, nm	3 – 20	2 – 10	irregular shape, ca. 10 nm size,
Use	precursor for carbon aerogels		precursor for carbon aerogels
Color	dark red, transparent	colorless, transparent	dark brown

RF and MF aerogels consist of interlinked colloidal particles formed during sol – gel processing. The size and the number of the clusters strongly depends on the resorcinol/catalyst (R/C) ratio. Typical R/C values are in the range of 50 – 300. High catalyst concentrations (e.g., R/C = 50) result in particles with diameters of 3 – 5 nm, interconnected by broad particle necks (polymeric gels). If the R/C value is increased to 200 – 300, the particle diameter increases (11 – 14 nm) and they cross-link like pearls on a string (colloidal gels) [152]. No fractal character was found for organic RF aerogels in SAXS experiments [155]. Nanophase separation, both initiated and limited by cross-linking, was discussed as an alternative explanation for network formation [156].

2.2. Drying Methods

The evaporation of the liquid from a wet gel proceeds in a very complex way. In the first stage, the gel shrinks by the volume that was previously occupied by the liquid. The liquid flows from the interior of the gel body to its surface. If the network is compliant, as in alkoxide-derived gels, the gel deforms (Figure 7).

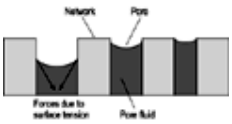

Figure 7. Representation of the contracting surface forces in pores of different size during drying. At the same pressure, the curvature of all menisci in the pores is the same. For this reason the larger pores are emptied first (after [32])

On shrinkage, OH groups on the inner surface approach each other and can irreversibly form M–O–M bridges. As drying proceeds, the network becomes increasingly stiffer, and the surface tension in the liquid rises correspondingly because the pore radii become smaller. The second stage of the drying process begins when the surface tension is no longer capable of deforming the network and the gel body becomes too stiff for further shrinkage. This point is called the critical point. (The critical point of the drying process should not be mistaken for the critical point at which a liquid changes into the supercritical state.) At this stage the tension in the gel becomes so large that the probability of cracking is highest. Although the liquid/gas interface retreats into the gel body, a contiguous funicular liquid film remains on the pore walls, that is, most of the liquid still evaporates from the exterior surface of the gel body. In the third stage of drying, this film is ruptured. Eventually, there is only liquid in isolated pockets, which can leave the network only by diffusion via the gas phase [17], [157].

Two processes are important for the collapse of the network: First, the slower shrinkage of the network in the interior of the gel body results in a pressure gradient which causes cracks. Second, larger pores are emptied faster than smaller pores during drying, so that if pores of different radii are present, the meniscus of the liquid drops faster in larger pores (Fig. 7). The walls between pores of different size are therefore subjected to nonuniform stress and therefore crack.

The parameters influencing these capillary forces are well investigated, and strategies were developed to avoid cracking. Important options are, inter alia, the pore size (larger pores are advantageous, because the capillary force is inversely proportional to the pore radius) [158], [159], aging (the network becomes stiffer), and addition of tensides (reduction of the surface energy leads to a decrease in capillary forces), organic templates, or drying control chemical additives (DCCA) [160-164]. Crack-free xerogels can be obtained by exact control of these parameters, although large shrinkage cannot be avoided.

This is not sufficient for the production of aerogels. Methods must be applied which both conserve the pore structure of the wet gels and avoid shrinkage.

2.2.1. *Supercritical Drying*

In supercritical drying, the solvent is put into the supercritical state, so that the liquid/gas interface in the pores disappears.

The wet gel is placed in an autoclave and covered with additional solvent. Drying of the samples, which would lead to the formation of cracks, is thus avoided. After closing the autoclave, the temperature is slowly raised, resulting in a pressure increase (path A in Fig. 8). Both the temperature and the pressure are adjusted to values above the critical point of the solvent (p_c, T_c) and kept at these values for a certain period. This ensures that the autoclave is completely filled with the supercritical fluid. The fluid is then slowly vented at constant temperature, while the pressure drops. When ambient pressure is reached, the vessel is cooled to room temperature and then opened. Drying is often performed with the vessel pre-pressurized by nitrogen to avoid evaporation of the solvent (path B in Fig. 8) [165]. The phase boundary between the liquid and the gas need not be crossed during drying.

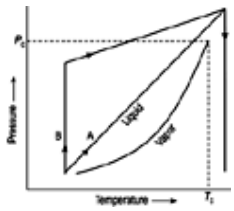

Figure 8. Schematic representation of the principle of supercritical drying. At the critical point [T_c, p_c] the density of the liquid and the gas are equal. Supercritical drying can be performed along path A or B

2.2.1.1. *Drying in Organic Solvents*

The pore liquid, which is usually an alcohol or acetone due to the preparation conditions, is used as the supercritical fluid. About 250 °C and 7 MPa are needed to put these solvents in the supercritical state. Problems arise from the combination of high temperature, high pressure, and the flammability of the solvents [166].

Furthermore, rearrangement reactions in the gel network are highly probable because of the high temperature. For example, when aerogels are dried with alcohols, a modification of the surface may occur by conversion of M–OH to M–OR groups, or processes such as those occurring during aging and ripening of the gels are accelerated. Small pores are thus filled, and particle necks are strengthened. As a consequence, the resulting aerogels have a lower specific surface area, a narrow pore radius distribution (elimination of microporosity), and a stiffer network [167-170]. Organic groups may be destroyed during drying [171]. Phase separation and loss of stoichiometry (SiO_2 – B_2O_3, SiO_2 – Bi_2O_5 [172]) or crystallization (SiO_2 – TiO_2 [173]) may occur in multicomponent gels [5], [174].

2.2.1.2. *Supercritical Drying with CO_2*

An alternative to drying in organic solvents is the use of liquid carbon dioxide [175-177]. This method has the advantage of a very low critical temperature at a moderate critical pressure. However, time-consuming solvent exchange is necessary before drying. Another requirement is the miscibility of the pore liquid with carbon dioxide. For example, water and CO_2 are immiscible; therefore, an intermediate solvent exchange (e.g., water against acetone) is necessary [103]. Structural changes, as observed during supercritical drying with alcohols, only take place to a minor extent with CO_2. Therefore, the network of the wet gel can be conserved almost unchanged.

A modification of this method and its application in a continuous plant in which CO_2 is recycled is described in [178]. Heating and cooling of the autoclave is avoided by exchanging the solvent with supercritical CO_2.

Deliberate structural changes were induced in SiO_2/TiO_2, ZrO_2, and Nb_2O_5 aerogels by variation of the temperature above T_c during supercritical drying with CO_2. For example, the ratio between crystalline and amorphous fractions can thus be influenced, which is important for catalytic applications [179].

2.2.2. ***Freeze Drying***

Another possibility to avoid phase boundaries between the liquid and the gas phase during drying is freeze drying. The pore liquid is frozen and then sublimed in vacuo. There were some attempts to use this method for the production of aerogels [180-182]. However, the aging period must be extended to stabilize the gel network, the solvent must be replaced by one with a low expansion coefficient and a high sublimation pressure, and low freezing temperatures are attained by addition of salts. Another disadvantage is that the network may be destroyed by crystallization of the solvent in the pores. Cryogels are therefore only obtained as powders.

2.2.3. ***Ambient Pressure Drying***

Supercritical drying is the most expensive and riskiest part of the preparation of aerogels and must be dispensed with if aerogels are to become of interest for large-scale commercial applications. Therefore, the interest in other possibilities for exchanging the pore liquid in the gels by air is very high.

The capillary forces exerted by the meniscus of the pore liquid and the pressure gradient due to the shrinkage of the network are the main reasons for the collapse of the filigree structure. Drying of the gels under ambient conditions (also known as subcritical drying) requires strengthening of the network to avoid its collapse (irreversible shrinkage). Additionally or alternatively, the contact angle between the pore liquid and the pore walls must be influenced by deliberate modification of the inner surface and variation of the solvent to minimize capillary forces.

Ambient-pressure drying involves a series of solvent-exchange processes and modification of the inner surface [183], [184]. The water/alcohol mixture in the pores of the gel is first replaced by a water-free solvent, and the Si-OH groups at the surface are silylated, for example, with trimethylchlorosilane. The reactivity of the gel surface is thus reduced, and it is also hydrophobized. The actual drying process is performed after another solvent exchange. The gel shrinks strongly

during evaporation of the solvent from the pores, as expected. However, no irreversible narrowing of the pores by formation of Si–O–Si bonds is possible because of the silylation. Therefore, the gel expands to nearly its original size after reaching the critical point (spring-back effect). One of the prerequisites for this procedure is the strengthening of the network by exact control of the aging conditions. The network of the modified gels must be stable enough to tolerate a reversible shrinkage of up to 28 % of its original volume [185].

Similar good results were obtained by another strategy [186], [187]. The strength and stiffness of the network was drastically increased by aging the wet gels in solutions of tetraalkoxysilanes in aqueous alcohols. Shrinkage during drying was thus completely avoided. The monomers added during aging mainly condense in the smallest pores and at the particle necks. The microporosity of the gels is thus lost. The gels can also be dried at ambient pressure and 20 – 180 °C because of the stiffer network.

With the ambient-pressure drying technique for the preparation of aerogels, a technically much simpler and therefore less expensive method of preparation is available, which makes the industrial application of aerogels economically interesting [188]. A similar procedure based on solvent-exchange processes and drying under ambient conditions was successfully invented for the drying process of organic RF aerogels [189], [190].

Aerogels can be prepared as monoliths, granulates, or powders. However, many applications require porous films (coatings), for which supercritical drying is not a trivial issue. Nevertheless, aerogel films were synthesized in some laboratories [191], [192]. The possibilities provided by ambient-pressure drying eliminate the principal obstacles for preparing highly porous films [193].

2.3. Modification of Aerogels after Drying

Chemical modifications are still possible after drying. However, post-synthesis treatment with a liquid phase is problematic because wet gels are formed again.

Aerogels can be doped with metals by impregnation with a metal salt. The metal salt has to be introduced in alcoholic solution, and then the solvent has to be removed again in another supercritical drying step to prevent destruction of the aerogel network [121]. Another, more general possibility for the subsequent modification of aerogels is their reaction with gaseous compounds. For example, gaseous dichlorodimethylsilane or other silylating agents were employed to hydrophobize the aerogel permanently [194]. Deposition of carbon by decomposition of gases such as acetylene, methane, propane, xylene, or furfuryl alcohol was used to improve the thermal properties of aerogels [195-202], [203].

The chemical vapor infiltration (CVI) method was employed to prepare SiO_2 – Si nanocomposites by reaction with SiH_4 or $HSiCl_3$ and for the deposition of iron particles from ferrocene [204]. Reaction of SiO_2 aerogels with $[W(CO)_6]$ and subsequent thermal decomposition of the metal carbonyl in the aerogel resulted in metallic tungsten particles. WN or WS_2 particles in aerogels were formed by aftertreatment of the metal-doped aerogels with ammonia or sulfur. SiO_2 aerogels were analogously treated with gaseous $[Fe(CO)_5]$. In an oxidizing atmosphere, amorphous iron oxide/SiO_2 aerogels were obtained, which in subsequent reactions were transformed into Fe_3O_4/SiO_2 (by reaction with additional $[Fe(CO)_5]$) or to Fe_9S_{10}/SiO_2 aerogels (by reaction with H_2S), for example [205].

Organic compounds can also be sublimed into the network via the gas phase, as shown for a porphyrin derivative [206].

Beside the treatment of aerogels with gaseous compounds, pyrolysis of organic groups which are part of the aerogel structure is another possibility for modifying aerogels. The coating of the inner aerogel surface with carbon structures was thus achieved [203].

Carbon aerogels are of particular interest because they are electrically conducting [207-210]. They can be made by pyrolysis of resorcinol/formaldehyde or phenol/furfural aerogels under an inert atmosphere (1050 °C, N_2). Their structure and morphology basically corresponds to that of the precursor gels, that is, they are colloidal or polymeric (Table 8). The colloidal gels have spherical particles with diameters of 12 – 15 nm which are loosely connected by graphitic ribbons. The polymeric gels have smaller particles (7 – 9 nm) connected by broad particle necks. The precursor to catalyst ratio during the preparation of the organic gels and the pyrolysis temperature have the largest influence on the properties of carbon aerogels.

Table 8. Typical structural data of carbon aerogels

Skeletal density, g/cm^3	2.06
Density, g/cm^3	0.05 – 0.80
Specific surface area, m^2/g	600 – 800
Pore size, nm	< 50
Particle morphology, nm	3 – 20, amorphous
Appearance	black, opaque

[Top of Page]

3. Properties

3.1. Structural Properties

The microstructure of aerogels strongly depends on the preparation conditions and the choice of precursors. Therefore, each aerogel has its own structural characteristics (Chap. Synthesis).

The gathering of information on the structure of aerogels requires methods which cover the length scale from the lower nanometer range (structure of the primary particles) to the micrometer range (linking of the particles, etc.) [211]. No single method can provide the complete information. Some common methods for the structure determination of aerogels are mentioned in Table 9.

Table 9. Typical methods for the structural characterization of aerogels and the structural information available therefrom

Method	Structural information	Remarks/problems
N_2 adsorption and desorption (BET) [212]	specific surface area	only pores which are accessible to N_2 are detected
	pore radius distributions [213]	structural changes caused by the N_2 pressure during the measurements results in falsification of the values for the pore radii [214]
	adsorbent/ adsorbate interactions	
Helium pyknometry [215]	skeletal density	
Mercury porosimetry [216]	pore volume	compression of the network by the applied pressure instead of penetration of Hg into the matrix [217], [218]
Electron microscopy SEM (scanning), TEM (transmission), STEM (scanning transmission electron microscopy)	particle size	sintering processes due to the high energy of the electron beam
	form and arrangement of the particles (morphology)	
Light scattering		
Small angle X-ray scattering (SAXS)	particle and cluster sizes	inverse relation of the scattering vector and the structural dimension
	fractal dimensions	
Small angle neutron scattering (SANS)	surface area	information from the Ångstrom to the micrometer range
Nuclear magnetic resonance (NMR)	structure evolution from the molecular precursors to the gel stage [219], [220]	
	pore diameters or fractal dimensions [221], [222]	
Atomic force microscopy (AFM)	atomic structure at the surface [223]	

3.2. Physical and Chemical Properties

3.2.1. *Optical Properties*

The optical properties of SiO_2 aerogels and derived inorganic/organic hybrid aerogels vary between transparent and translucent, depending on the preparation conditions. This means that the structure-forming units are smaller than the wavelength of visible light. However, inhomogeneities in the network in the nanometer range result in Rayleigh scattering. Therefore, aerogels appear yellowish under light and bluish against a dark background. The portion of Rayleigh scattering correlates directly with the homogeneity of the gel network in the nanometer range. Objects that are viewed through an aerogel tile appear slightly blurred. Inhomogeneities in the micrometer range at the exterior surface of the aerogel body are responsible for this phenomenon [224].

For most applications a high optical transmission is desirable; therefore, numerous attempts were made to influence the optical properties. Improvements were achieved by physical approaches such as desorption of water, different drying procedures, and adjusting the synthesis conditions [4], [177], [225-232].

3.2.2. *Mechanical and Acoustic Properties*

Mechanical properties have many aspects. Among others, rigidity, brittleness, and dimensional stability, as well as the method of investigation, have to be considered. Although the rigidity of aerogels is low because of their structure and the brittleness of the SiO_2 particles, it is sufficient for most applications. The compressive strength of monolithic SiO_2 aerogels is in the range 0.15 – 0.30 N/mm^2, depending on the density, with an elastic compression of about 2 – 4 %. The rigidity in vacuum is distinctly higher. The tensile strength is about 0.020 N/mm^2 [233].

One of the decisive parameters for the mechanical properties is the interconnection of the network in the nanometer range, which strongly depends on the preparation conditions. For example, aerogels prepared under acidic or neutral conditions are about twice as stiff as the corresponding samples prepared under base-catalyzed conditions [234]. Investigations of organically modified aerogels showed that the elastic properties of SiO_2 aerogels can also be improved by using chemically modified precursors. Since the longitudinal sound velocity is related to one of the elastic constants (which correlates with Young's modulus) and the density of the aerogel, ultrasound measurements were used for these investigations. The reason for the improved elastic properties of hybrid aerogels is probably that the presence of alkyl or aryl groups on the surface of the secondary particles impairs stiffening of the particle necks by Ostwald ripening [141], [235].

The particular network structure of aerogels results in impressive acoustic properties which, as mentioned above, correlate with the elastic properties [236]. The sound velocities in SiO_2 aerogels of 100 – 300 m/s are among the lowest for inorganic solids (which is noteworthy when one considers that sound velocities of 5000 m/s were measured in quartz glass) [237], [238]. SiO_2 aerogels have the lowest acoustic impedance of all solid materials ($Z = 10^4 – 10^5$ kg m^{-2} s^{-1}).

3.2.3. *Thermal Conductivity*

The thermal conductivity of aerogels is made up of three contributions: thermal transport by the gas phase, by the solid phase, and by radiation. Figure 9 schematically shows the contribution of these components as a function of bulk density.

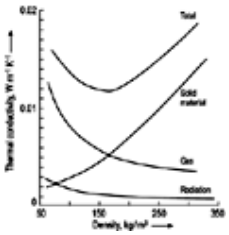

Figure 9. Composition of the thermal conductivity of aerogels: gas, solid and radiation transport as a function of the density of the material (reprinted from [239] with permission of Spektrum der Wissenschaft Verlag)

The solid conductivity clearly increases with increasing density, whereas the reverse is true for the gas and radiative transport. The minimum for the total thermal conductivity is at a density of about 0.150 g/cm^3 [239]. Heat transport by the gas phase can be avoided by evacuation of the system. For example, an evacuated SiO_2 aerogel at about 300 K shows a thermal conductivity of only 0.010 W m^{-1} K^{-1}, compared to 0.020 W m^{-1} K^{-1} in air [240]. This is not a specific property of SiO_2 aerogels. Thermal conductivities as low as 0.012 W m^{-1} K^{-1} were measured for nonevacuated organic RF aerogels [241].

Another possibility for lowering the thermal conductivity of SiO_2 aerogels is to reduce the radiative heat transport. This is more important for high-temperature applications, because the radiation maximum is then in a wavelength range in which SiO_2 does not sufficiently absorb (3 – 5 µm). Reduction of thermal conductivity at these temperatures is possible by doping the aerogels with infrared opacifiers, such as carbon black or TiO_2. These compounds can be mixed in a highly disperse form with the precursor solution during sol – gel processing or with the sol [241]. Another possibility is the generation of carbon in the aerogel by pyrolysis of organic gases or covalently bonded organic groups of hybrid aerogels (see Sections Inorganic/Organic Hybrid Aerogels and Modification of Aerogels after Drying). The latter method leads to a particularly efficient decrease of the thermal conductivity (0.012 W m^{-1} K^{-1} in air at 300 K), because the aerogel skeleton is partially covered with nanosized carbon structures [242].

3.2.4. *Hydrophobicity*

A major drawback of unmodified SiO_2 aerogels, in particular for technical use, is their long-term instability under a humid atmosphere. Owing to the large number of silanol groups on the inner surface of the material, adsorption and capillary condensation of water and eventually cracking of the gel body by the resulting capillary forces results. Approaches to permanently hydrophobize aerogels were therefore investigated early on. The possibility of treating an aerogel subsequently with a silylating agent has the disadvantage of complicated processing. This kind of modification is only possible via the gas phase (see Section Modification of Aerogels after Drying).

Hydrophobation at the alcogel stage is time-consuming, since the diffusion rate of the silylating reagent into the pores is a limiting factor. However, this process is a key step in the prevailing ambient-pressure drying of aerogels (see Section Ambient Pressure Drying). Aerogels prepared by this method are inherently hydrophobic.

[Top of Page]

4. Applications

SiO_2 aerogels have the highest potential for applications. They are not only the best-investigated with regard to their preparation, structure, and properties, but can also be prepared with lower costs than other aerogels. There are already commercial applications. Nonsilicate aerogels (see Section Metal Oxide Aerogels) are playing an increasingly important role, particularly in the search for new catalysts. In the case of organic aerogels, mainly their pyrolysis products (carbon aerogels) offer interesting new applications [243].

4.1. Çerenkov Detectors

Çerenkov radiation is emitted by charged particles such as pions, muons, or protons when they exceed the speed of light in a medium. A cone-shaped electromagnetic shock wave is formed. The velocity of the particles can be calculated from the cone angle. The speed of light depends on the refractive index of the medium. SiO_2 aerogels with refractive indices between 1.007 and 1.024 offer a range not covered by the previously used compressed gases or liquids. As solid materials, aerogels additionally allow easier construction of the detectors and are therefore ideal materials for this application [239], [244]. They are now installed in several Çerenkov detectors worldwide, and the amount of aerogel tiles produced for this application is impressive. For example, at the Deutsches Elektronen-Synchroton (DESY) in Hamburg, a detector was supplied with about 1700 L of silica aerogel.

4.2. Thermal Insulation

Aerogels have the lowest thermal conductivities of all solids (Figure 10). They are among the best thermal insulation materials. Additional advantages are nonflammability and transparency.

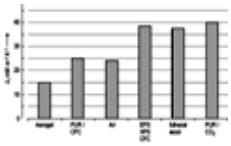

Figure 10. Comparison of the thermal insulation properties of some commercially available insulating materials (PUR = polyurethane foam; CFC = chlorofluorocarbons; EPS, XPS = expanded and extruded polystyrene). (Reprinted from [188] with permission of Hoechst AG)

However, a major problem in the insulation of windows is the production of large aerogel panels [245]. Conventional windows cannot be equipped with the commonly prepared granulate, which is not transparent due to the roughness of its surface; it can only be used for frosted glass windows (for example, bathroom, staircase, or ceiling windows). The scattering of the sunlight in the granular layer is here rather advantageous, because diffuse lighting of the rooms is achieved.

In addition to transparent thermal insulation, applications for a variety of other insulation problems are conceivable for which transparency of the insulating material is not necessary, for example, in cooling or heating systems, high-temperature batteries, insulation for piping, heat and cold storage appliances, automotive exhaust pipes, transport vehicles and vessels, etc. Thin aerogel films can even be used for coatings on IR detectors to shield against the heat radiating from the substrate [7-11], [246-248].

Aerogels can also be utilized for passive exploitation of solar energy, for example for paneling of house walls (Figure 11) or for coating solar-energy collectors [249-253]. A layer of the transparent aerogel allows the penetration of the sun′s radiation to the blackened wall, but not the escape of the heat generated there [254].

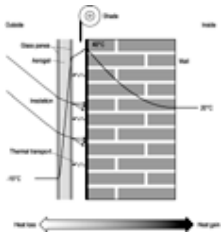

Figure 11. Concept for energy saving by passive use of solar energy (reprinted from [239] with permission of Spektrum der Wissenschaft Verlag)

4.3. Catalysis

All the advantages of sol – gel processing for preparing catalysts or catalyst supports, especially the controllable dispersion of the active component and the possibility of deliberately influencing the microstructure, also apply for aerogels. Additionally, typical structural properties predestine them for these applications: the extremely high inner surface area (which is readily accessible because of the open porosity), the large pore volume, excellent temperature stability, and the high selectivities in catalytic processes. Most oxides relevant for catalytic applications can be prepared as aerogels, including binary and ternary compositions, mixtures of metal oxides, or metal particles on oxide supports. Furthermore, aerogels can be prepared as powders, granulates, or monoliths, either amorphous or with crystalline fractions.

Despite these obvious advantages, the use of aerogels in catalysis has only been increasingly investigated since 1974. Considerable progress was achieved in solving problems originating from the low density and the low thermal conductivity, that is, pressure drops and poor heat dissipation in the catalyst bed. For example, heat-dissipating materials can be added, or the aerogels can be deposited during their preparation (i.e., in the autoclave) on a support (such as Raschig rings, metal tubes, cordierite honeycombs, wire mesh screens, boiling stones, and others) [74].

It is impossible to list all the investigated systems here. They are summarized in specialist reviews [3], [5]. The following selection is intended to give an impression of the numerous possibilities: nitroxidation of hydrocarbons to nitriles (NiO – Al_2O_3 with SiO_2 and MgO, PbO – ZrO_2) [101], [111] hydrogenation of nitrobenzene to aniline and of toluene to methylcyclohexane (Ni – SiO_2, Pd – Al_2O_3) [125], hydrogenation and hydrogenolysis after hydrogen "spillover" activation (ZrO_2, SiO_2, MgO, Al_2O_3) [255], isomerizations (Nb_2O_5, Nb_2O_5 – SiO_2) [81], Fischer – Tropsch syntheses (Fe_2O_3 – Al_2O_3, Fe_2O_3 – SiO_2) [98], polymerization of ethene or propene ($TiCl_4$ – Al_2O_3) [256], deNO_x reactions (Fe_2O_3 – Cr_2O_3 – Al_2O_3, Pt – C) [151], [257], catalysis for combustion of car exhaust (Pd – Al_2O_3, Pt – Al_2O_3, Pd – Al_2O_3 – SiO_2) [123], [258], and fluorinations (Cr_2O_3) [75].

4.4. Application as Storage Media

The high porosity and very large surface area of silica aerogels can also be utilized for applications as gas filters, absorbing media for desiccation and (nuclear) waste containment, encapsulation media, and hydrogen fuel storage [2], [259-264].

Partially sintered aerogels can resist the tensions of a gas/liquid interface because their texture is strengthened during sintering. They can therefore be used for the storage, thickening, or transport of liquids, for example rocket fuels; in this case the low weight of aerogels is particularly advantageous [2]. Another potential application of aerogels relates to energetic materials. Traditional manufacturing involves processing of granular solids. Their performance strongly depends on the particle size distribution, surface area, homogeneity, etc. These issues could be improved by preparing detonators from aerogels [265].

4.5. Electrical and Electronic Applications

Carbon aerogels are very well suited as electrode materials for batteries and capacitors because of their low electric resistance (< 40 mΩ cm), controllable pore radius distribution (1 – 50 nm), and high specific surface area. When a voltage is applied, a thin polarization layer of high capacity is formed on the large surface. More charge can be stored than with conventional capacitors, because the surface area per unit volume of aerogels is very high. Compared to other materials (carbon powder or fibers), aerogels have the major advantage that they can be prepared as monoliths [207-210]. One of the most promising applications in this area is a cost-effective purification process using carbon aerogel electrodes for a variety of applications ranging from extracting harmful contaminants from industrial wastewater to desalinating sea water [266], [267].

Metal oxide and organic aerogels are excellent materials for low-dielectric applications. While bulk aerogels can be used for microwave electronics and high voltage insulators, thin aerogel films can compete with existing technologies in microelectronics. Here, the low dielectric constant of aerogels (< 2.0) plays a particularly important role. They offer important progress in the velocity of electric circuits. In addition, the dielectric constant is easily controlled because it is directly related to the porosity and hence the density of the materials.

Various metal oxide aerogels have been synthesized which exhibit superconducting, thermoelectric, or piezoelectric behavior [268-270].

4.6. Acoustic and Mechanical Applications

Aerogels have very low acoustic impedances and can therefore be used for matching the high acoustic impedance of piezo ceramics to the low impedance of air. The efficiency of these devices is considerably increased by decreasing the reflection losses at this interface [271], [272].

Aerogels have also been proposed as shock- and sound-absorbing materials (anechoic chambers) [272], [273]. Another interesting application is the use as sound-insulation material. The acoustic insulation performance of silica aerogel composite materials in the low frequency range is very good [274].

4.7. Optical Applications

Aerogels can be sintered at low temperatures and can therefore be processed into extremely pure and totally homogeneous glasses [275-280]. The inner surface area and the porosity decrease during sintering. The pore structure can therefore be modified in a controlled way by partial sintering. This was used, for example, for producing gas filters with pores in the range of 20 – 100 nm [281]. Doping of aerogels with dyes or rare earth elements allows the fabrication of glass for applications such as lasers and radioluminescent light sources [282-286].

Aerogel films can be exploited as optical coatings for solar cells because of their low refractive index (aerogel films prepared for this purpose are subsequently attached to the surface of the solar cell) [287]. The glass layers with high refractive indices otherwise used can thereby be replaced by aerogels. More light reaches the active surface because of the lower Fresnel scattering losses. These films have also been applied to the outside of fiber optics to improve the light collection and propagation efficiency.

4.8. Other Applications

There are numerous potential applications of aerogels. SiO_2 aerogels are suitable as ecologically acceptable insecticides against pests in grain and seeds, or those on the ground, such as mites, ticks, cockroaches, and silverfish [288-291]. As a chemically inert dust with small particles and a high specific surface area, aerogels can remove the protecting lipid layer of insects by abrasion or adsorption. They are then killed by rapid loss of body liquid.

Low-density silica aerogels can be used to study the superfluid transition of ^4He and phase separation of ^3He – ^4He mixtures [238].

SiO_2 aerogels are one of the standard media for collecting cosmic dust [292], and they are attached to the outside of spacecraft [293]. Aerogels have the advantage that they are extremely light with a unique mesostructure which is also transparent. This enables a soft landing of very fast extraterrestrial particles impinging with a velocity of more than 3 km/s, and also allows their subsequent investigation by optical methods.

[Top of Page]

5. Future Perspectives

The unique optical, thermal, acoustic, and mechanical properties of aerogels originate from the combination of a solid matrix (with a certain chemical composition) and nanosized pores filled with air. There is a direct connection between the chemistry of the sol – gel process and the structure of the gels, and between the structure and the properties of the aerogels. The choice of the precursors and the chemical reaction parameters already determines the physical properties of the final materials.

The possibilities originating from ambient-pressure drying will provide strong impetus to preparative chemical research, because aerogels can now be prepared with standard laboratory equipment. There is definitely room for more extensive chemical modifications of the property profile of aerogels.

The new drying techniques will probably make technical preparation much cheaper and will thus make aerogels more competitive. Apart from special applications in which material costs only play a minor role, the rather high price of supercritically dried aerogels mitigated against a broader range of applications. The most important area for the application of aerogels are all kinds of thermal insulation. There is no doubt about the physical and ecological advantages (nontoxic, nonflammable, easy to dispose of) of SiO_2 aerogels compared with most other materials on the market.

[Top of Page]

References

General References

1. R. J. Ayen, P. A. Iacobucci, *Rev. Chem. Eng.* **5** (1988) 157.
2. H. D. Gesser, P. C. Goswami, *Chem. Rev.* **89** (1989) 765. Links
3. G. M. Pajonk, *Appl. Catal.* **72** (1991) 217; *Catal. Today* **35** (1997) 319. Links
4. J. Fricke, A. Emmerling, *Struct. Bonding* **77** (1992) 37.
5. M. Schneider, A. Baiker, *Catal. Rev. – Sci. Eng.* **37** (1995) 515; *Catal. Today* **35** (1997) 339. Links
6. N. Hüsing, U. Schubert, *Angew. Chem. Int. Ed.* **37** (1998) 22. Links
7. J. Fricke (ed.): *Aerogels (Springer Proc. Phys. 6)*, Springer, Berlin 1986.
8. “Proc. Int. Symp. Aerogels 2nd,” Rev. Phys. Appl., Colloque C4, Les Editions de Physique, Les Ulis Cedex 1989
9. “Proceedings of the 3rd Int. Symp. Aerogels,” *J. Non-Cryst. Solids* **145** (1992).
10. “Proceedings of the 4th Int. Symp. Aerogels,” *J. Non-Cryst. Solids* **186** (1995).
11. “Proceedings of the 5th Int. Symp. Aerogels,” *J. Non-Cryst. Solids* **225** (1998).
12. “Proceedings of the 6th Int. Symp. Aerogels,” *J. Non-Cryst. Solids* **285** (2001).
13. “Proceedings of the 7th Int. Symp. Aerogels,” *J. Non-Cryst. Solids* **350** (2004).
14. F. Schüth, K. S. W. Sing, J. Weitkamp (eds.): *Handbook of Porous Solids*, Wiley-VCH, Weinheim, 2002, Vol. 3., pp. 2014.

Specific References

15. K. S. W. Sing *et al.*, *Pure Appl. Chem.* **57** (1985) 603. Links
16. D. W. Hua, J. Anderson, J. Di Gregorio, D. M. Smith, G. Beaucage in [10], p. 142.
17. C. J. Brinker, G. W. Scherer: *Sol – Gel – Science*, Academic Press, New York 1990.
18. Y. A. Attia (ed.): *Sol – Gel Processing and Applications*, Plenum, New York 1994.
19. Symposium series *Better Ceramics Through Chemistry:* a) Mat. Res. Soc. Symp. Proc. **32** (1984); b) **73** (1986); c) **121** (1988); d) **180** (1990); e) **271** (1992); f) **346** (1994); g) **435** (1996).
20. J. D. Mackenzie, D. R. Ulrich (eds.): *Ultrastructure Processing of Advanced Ceramics*, J. Wiley & Sons, New York 1988.
21. D. R. Uhlmann, D. R. Ulrich (eds.): *Ultrastructure Processing of Advanced Materials*, J. Wiley & Sons, New York 1992.
22. L. L. Hench, J. K. West (eds.): *Chemical Processing of Advanced Materials*, J. Wiley & Sons, New York 1992.
23. H. Dislich in L. C. Klein (ed.): *Sol – Gel Technology of Thin Films, Fibers, Preforms, Electronic and Speciality Shapes*, Noyes, Park Ridge, New York 1988, p. 50.
24. L. L. Hench, J. K. West, *Chem. Rev.* **90** (1990) 33. Links
25. D. W. Schaefer, *Science* **243** (1989) 1023. Links
26. G. Orcel, R. W. Gould, L. L. Hench in [19, b], p. 289.
27. T. Lours, *J. Non-Cryst. Solids* **100** (1988) 207. Links
28. A. Hasmy, R. Jullien in [10], p. 442.
29. P. I. Pohl, J.-L. Faulon, D.-M. Smith in [10], p. 449.
30. E. A. Chandler, D. F. Calef in [10], p. 456.

31. D. W. Schaefer, K. D. Keefer, in L. Pietronero, E. Toscatti (eds.): *Fractals in Physics*, North-Holland, Amsterdam 1986, p. 39.
32. D. W. Schaefer, *MRS Bull.* **19** (1994) no. 4, 49. Links
33. K. D. Keefer, D. W. Schaefer, *Phys. Rev. Lett.* **53** (1984) 1383.
34. K. D. Keefer in [19, a], p. 1.
35. J. Livage, M. Henry, C. Sanchez, *Prog. Solid State Chem.* **18** (1988) 259.
36. K. D. Keefer, D. W. Schaefer, *Phys. Rev. Lett.* **56** (1986) 2199, 2376.
37. T. Heinrich, U. Klett, J. Fricke, *J. Porous Mat.* **1** (1995) 7. Links
38. C. J. Brinker, K. D. Keefer, D. W. Schaefer, C. S. Ashley, *J. Non-Cryst. Solids.* **48** (1982) 47. Links
39. T. M. Tillotson, L. W. Hrubesh, I. M. Thomas in [19, c], p. 658.
40. T. M. Tillotson, L. W. Hrubesh, in [9], p. 44.
41. G. M. Pajonk *et al.* in [10], p. 1.
42. G. M. Pajonk *et al.* in [18], p. 275.
43. Hoechst, DE 4 316 540 A 1, 1993 (R.-M. Jansen, B. Kessler, J. Wonner, A. Zimmermann).
44. Hoechst, EP 690 023 A 2, 1993 (R.-M. Jansen, A. Zimmermann).
45. Henkel, EP 409 083, 1991 (H. Dolhaine, J. Wolf-Leistekow).
46. G. Herrmann, R. Iden, M. Mielke, F. Teich, B. Ziegler in [10], p. 380.
47. C. A. M. Mulder, J. G. van Lierop in [7], p. 68.
48. C. Sanchez, F. Ribot, *New J. Chem.* **18** (1994) 1007. Links
49. U. Schubert, N. Hüsing, A. Lorenz, *Chem. Mater.* **7** (1995) 2010. Links
50. S. J. Teichner, G. A. Nicolaon, M. A. Vicarini, G. E. E. Gardes, *Adv. Colloid Interface Sci.* **5** (1976) 245. Links
51. S. J. Teichner in [7], p. 22.
52. G. Tournier, M. Lacroix-Repellin, G. M. Pajonk, S. J. Teichner in B. Delmon, P. Grange, P. A. Jacobs, G. Poncelet (eds.): *Studies in Surface Science and Catalysis*, vol. **31**, Elsevier, Amsterdam 1987, p. 333.
53. Y. Mizushima, M. Hori, *J. Non-Cryst. Solids* **170** (1994) 215. Links
54. Y. Mizushima, M. Hori, *J. Non-Cryst. Solids* **167** (1994) 1. Links
55. M. B. Vicarini, G. A. Nicolaon, S. J. Teichner, *Bull. Soc. Chim. Fr.* (1969) 1466.
56. S. J. Teichner, *Rev. Phys. Appl.* **24** (1989) 1.
57. C. J. Brodsky, E. I. Ko, *J. Mater. Chem.* **4** (1994) 651. Links
58. L. K. Campbell, B. K. Na, E. I. Ko, *Chem. Mater.* **4** (1992) 1329. Links
59. D. A. Ward, E. I. Ko, *Chem. Mater.* **5** (1993) 956. Links
60. M. Schneider, A. Baiker, *J. Mater. Chem.* **2** (1992) 587. Links
61. H. Vesteghem, T. Jacoon, A. Lecomte in [8], p. 59.
62. Z. Zhu, M. Tomkiewicz in [19, f], p. 751.
63. Z. Zhu, Y. Tsung, M. Tomkiewicz, *J. Phys. Chem.* **99** (1995) 15945.
64. Z. Zhu, M. Lin, G. Dagan, M. Tomkiewicz, *J. Phys. Chem.* **99** (1995) 15950.
65. A. Benedetti *et al.*, *J. Appl. Cryst.* **26** (1993) 717.
66. Y. W. Zeng, P. Riello, A. Benedetti, G. Fagherazzi, *J. Non-Cryst. Solids* **185** (1995) 78. Links
67. F. Chaput, B. Dunn, P. Fugua, K. Salloux, *J. Non-Cryst. Solids* **188** (1995) 11. Links
68. D. R. Rolison, B. Dunn, *J. Mater. Chem.* **11** (2001) 963. Links
69. H. Hiroshima, K. Sudoh in [19, e], p. 573.
70. H. Hiroshima, K. Sudoh in [9], p. 51.
71. K. Sudoh, H. Hiroshima, *J. Non-Cryst. Solids* **147 & 148** (1992) 386.
72. H. D. Foster, D. B. Keyes, *Ind. Eng. Chem.* **29** (1937) 1254.
73. K. Kearby, S. S. Kistler, S. Swann, *Ind. Eng. Chem.* **30** (1938) 1082.
74. J. N. Armor, E. J. Carlson, *Appl. Catal.* **19** (1985) 327.
75. Allied-Signal, US 4 828 818, 1989 (E. J. Carlson, J. N. Armor, W. J. Cunningham, A. M. Smith).
76. D. Bianchi, H. Batis-Landonsli, C. O. Bennett, G. M. Pajonk, P. Vergnon, S. J. Teichner, *Bull. Soc. Chim. Fr.* (1981) 345.
77. H. Zarouk, A. Ghorbel, G. M. Pajonk, S. J. Teichner, *Bull. Soc. Chim. Fr.* (1982) 71.
78. F. Blanchard, B. Pommier, J. P. Reymond, S. J. Teichner in G. Poncelet, P. Grange, P. A. Jacobs (eds.): *Studies in Surface Science and Catalysis*, vol. **16**, Elsevier, Amsterdam 1983, 395.
79. M. Astier *et al.* in B. Delmon, P. A. Jacobs, G. Poncelet (eds.): *Studies in Surface Science and Catalysis*, vol. **1**, Elsevier, Amsterdam 1976, p. 315.
80. M. Astier, A. Bertrand, S. J. Teichner, *Bull. Soc. Chim. Fr.* (1980) 191.
81. S. M. Maurer, E. I. Ko, *J. Catal.* **135** (1992) 125. Links S. M. Maurer, D. Ng, E. I. Ko, *Catal. Today* **16** (1993) 319. Links
82. E. C. DeCanio, V. P. Nero, E. I. Ko, *J. Catal.* **146** (1994) 317. Links
83. D. J. Suh, T.-J. Park, *Chem. Mater* **8** (1996) 509. Links
84. J. W. Long, K. E. Swider-Lyons, R. M. Stroud, D. R. Rolison, *Electrochem. Solid-State Lett.* **3** (2000) 453.

85. A. E. Gash, J. H. Satcher, Jr., R. L. Simpson in [13], p. 145.
86. J. L. Mohanan, S. L. Brock in [13], p. 1.
87. A. E. Gash, T. M. Tillotson, J. H. Satcher, J. F. Poco, L. W. Hrubesh, R. L. Simpson, *Chem. Mater.* **13** (2001) 999. Links
88. A. E. Gash, T. M. Tillotson, J. H. Satcher, L. W. Hrubesh, R. L. Simpson in [12], p. 22.
89. A. E. Gash, J. H. Satcher, R. L. Simpson, *Chem. Mater.* **15** (2003) 3268. Links
90. G. A. Nicolaon, S. J. Teichner, *Bull. Soc. Chim. Fr.* (1968) 1900.
91. G. E. E. Gardes, G. M. Pajonk, S. J. Teichner, *Bull. Soc. Chim. Fr.* (1976) 1321.
92. B. E. Yoldas, *J. Non-Cryst. Solids* **38** (1980) 81.
93. S. Yoda, K. Otake, Y. Takehayashi, T. Sugeta, T. Sato in [12], p. 8.
94. D. C. Dutoit, M. Schneider, A. Baiker, *J. Catal.* **153** (1995) 165. Links
95. R. Hutter, T. Mallat, A. Baiker, *J. Catal.* **153** (1995) 177. Links
96. F. Chaput, A. Lecomte, A. Dauger, J. P. Boilot in [8], p. 137.
97. T. Heinrich, F. Raether, H. Marsmann, *J. Non-Cryst. Solids* **168** (1994) 14. Links
98. F. Blanchard, J. P. Reymond, B. Pommier, S. J. Teichner, *J. Mol. Catal.* **17** (1982) 171.
99. R. Marinangelli, G. M. Pajonk, S. J. Teichner, *Proc. 7th Simposio Iberoamericano de Catalisis,* La Plata, Argentina 1980, p. 126.
100. S. Abouarnadasse, G. M. Pajonk, S. J. Teichner, *Actas Xe Simposio Iberoamericano de Catalisis,* vol. II, Merida, Venezuela 1986, p. 615.
101. G. M. Pajonk, T. Manzalji, *Catal. Lett.* **21** (1993) 361. Links
102. Y. Mizushima, M. Hori, *J. Mater. Res.* **9** (1994) 2272. Links
103. C. J. Brinker *et al.* in [7], p. 57.
104. P. Löbmann, W. Glaubitt, J. Gross, J. Fricke in [10], p. 59.
105. P. Löbmann, J. Gross, W. Glaubitt, J. Fricke, *J. Non-Cryst. Solids* **201** (1996) 66. Links
106. D. Demydov, K. J. Klabunde in [13], p. 165.
107. J. W. Long, M. S. Logan, C. P. Rhodes, E. E. Carpenter, R. M. Stroud, D. R. Rolison, *J. Am. Chem. Soc.* **126** (2004) 16 879. Links
108. J. W. Long, M. S. Logan, E. E. Carpenter, D. R. Rolison in [13], p. 182.
109. A. Sayari, A. Ghorbel, G. M. Pajonk, S. J. Teichner, *Bull. Soc. Chim. Fr.* (1981) 220.
110. J. N. Armor, E. J. Carlson, W. C. Conner, *React. Solids* **3** (1987) 155.
111. G. M. Pajonk in G. Centi, R. Trifiro (eds.): *Studies in Surface Science and Catalysis*, vol. **55**, Elsevier, Amsterdam 1990, p. 229.
112. G. Dagan, M. Tomkiewicz, *J. Non-Cryst. Solids* **175** (1994) 294. Links
113. B. Himmel, T. Gerber, H. Bürger, G. Holzhüter, A. Olbertz in [10], p. 149.
114. J. C. Pouxviel, J. P. Boilot in [20], p. 197.
115. T. Heinrich, F. Raether, W. Tappert, J. Fricke in [9], p. 55.
116. H. Hiroshima, M. Gengyo, C. Kojima, H. Imai in [10], p. 54.
117. G. M. Pajonk, M. B. Taghavi, S. J. Teichner, *Bull. Soc. Chim. Fr.* (1975) 983.
118. M. B. Taghavi, G. M. Pajonk, S. J. Teichner, *J. Colloid Interface Sci.* **71** (1979) 45.
119. M. Astier, A. Bertrand, S. J. Teichner, *Bull. Soc. Chim. Fr.* (1980) 218.
120. Solvay, EP 629 442 A 1, 1994 (B. Heinrichs, J.-P. Pirard, R. Pirard).
121. B. Pommier, S. J. Teichner in M. J. Phillips, M. Ternan, *Proc. 9th Int. Congr. Catal.*, **2** (1988) 610.
122. Y. Mizushima, M. Hori, *J. Mater. Res.* **10** (1995) 1424. Links
123. C. Hoang-Van, B. Pommier, R. Harivololona, P. Pichat in [9], p. 250.
124. M. Schneider, D. G. Duff, T. Mallat, M. Wildberger, A. Baiker, *J. Catal.* **147** (1994) 500. Links
125. J. N. Armor, E. J. Carlson, P. M. Zambri, *Appl. Catal.* **19** (1985) 339.
126. L. M. Hair *et al.* in [10], p. 168.
127. K. Balakrishnan, R. D. Gonzales, *J. Catal.* **144** (1993) 395. Links
128. F. Schwertfeger, W. Glaubitt, U. Schubert in [9], p. 85.
129. F. Schwertfeger, N. Hüsing, U. Schubert, *J. Sol – Gel Sci. Technol.* **2** (1994) 103.
130. N. Hüsing, F. Schwertfeger, U. Schubert, *J. Sol – Gel Sci. Technol.* **8** (1997) 807.
131. M. Bockhorst, K. Heinloth, G. M. Pajonk, R. Begag, E. Elaloui in [10], p. 388.
132. L. Zhu, Y. Li, J. Wang, J. Shen, *Chem. Phys. Lett.* **239** (1995) 393. Links
133. M. Power, B. Hosticka, E. Black, C. Daitch, P. Norris in [12], p. 303.
134. S. Maury, P. Buisson, A. C. Pierre, *Langmuir* **17** (2001) 295.
135. B. M. Novak, D. Auerbach, C. Verrier, *Chem. Mater.* **6** (1994) 282. Links
136. B. M. Novak, *Adv. Mater.* **5** (1993) 422. Links D. Avnir, L. C. Klein, D. Levy, U. Schubert, A. B. Wojcik in Z. Rappoport, Y. Apeloig (eds.): *The Chemistry of Organosilicon Compounds*, vol. **2**, "part. 2", J. Wiley & Sons, New York 1998, p. 2317.
137. B. Riegel, S. Plittersdorf, W. Kiefer, N. Hüsing, U. Schubert, *J. Mol. Struct.* **410 – 411** (1997) 157.
138. N. Hüsing, U. Schubert, B. Riegel, W. Kiefer in [19, g], p. 339.

139. N. Hüsing, U. Schubert, K. Misof, P. Fratzl, *Chem. Mater.* **10** (1998) 3024.
140. N. Hüsing, U. Schubert, *Mat. Res. Soc. Symp. Proc.* **576** 1999, 117.
141. U. Schubert, F. Schwertfeger, N. Hüsing, E. Seyfried in [19, f], p. 151.
142. W. Cao, A. J. Hunt in [19, f], p. 631.
143. N. Hüsing *et al.*, *Chem. Mater.* **11** (1999) no. 2, 451.
144. D. A. Loy *et al.* in [10], p. 44.
145. U.S. Department of Energy, US 4 873 218, 1989 (R. W. Pekala).
146. R. W. Pekala, *J. Mater Sci.* **24** (1989) 3221. Links
147. U.S. Department of Energy, US 4 997 804, 1991 (R. W. Pekala).
148. R. W. Pekala, in [21], p. 711.
149. R. W. Pekala, F. M. Kong, *Polym. Prep. (Am. Chem. Soc. Div. Polym. Chem.)* **30** (1989) 221.
150. R. L. Ward, R. W. Pekala, *Polym. Prep. (Am. Chem. Soc. Div. Polym. Chem.)* **31** (1990) 167.
151. R. W. Pekala, F. M. Kong in [8], p. 33.
152. G. C. Ruben, R. W. Pekala, T. M. Tillotson, L. W. Hrubesch, *J. Mater Sci.* **27** (1992) 4341. Links
153. R. W. Pekala, C. T. Alviso, X. Lu, J. Gross, J. Fricke, *J. Non-Cryst. Solids* **188** (1995) 34. Links
154. A. Rigacci, J. C. Marechal, M. Repoux, M. Moreno, P. Achard in [13], p. 372.
155. R. W. Pekala, D. W. Schaefer, *Macromolecules* **26** (1993) 5487. Links
156. D. W. Schaefer, *MRS Bull.* **24** (4) (1994) 14.
157. C. J. Brinker *et al.*, *J. Membrane Sci.* **94** (1994) 85. Links
158. P. J. Davis, C. J. Brinker, D. M. Smith, *J. Non-Cryst. Solids* **142** (1992) 189. Links
159. T. Mizuno, H. Nagata, S. Manabe, *J. Non-Cryst. Solids* **100** (1988) 236. Links
160. H. Chen, E. Ruckenstein, *J. Colloid Interface Sci.* **145** (1991) 581. Links
161. S. B. Chan, J. Jonas, *J. Non-Cryst. Solids* **126** (1990) 79. Links
162. C. Roger, M. J. Hampden-Smith, *J. Mater. Chem.* **2** (1992) 1111. Links
163. J. Zarzycki, M. Prassas, J. Phalippou, *J. Mater. Sci.* **17** (1982) 3371. Links
164. L. L. Hench, in L. L. Hench, D. R. Ulrich (eds.): *Science of Ceramic Chemical Processing*, J. Wiley & Sons, New York 1986, p. 52.
165. J. G. van Lierop, A. Huizing, W. C. P. M. Meermann, C. A. M. Mulder, *J. Non-Cryst. Solids* **82** (1986) 265. Links
166. S. Henning in [7], p. 38.
167. A. H. Boonstra, J. M. E. Baken, *J. Non-Cryst. Solids* **109** (1989) 1. Links
168. P. Wang, A. Emmerling, W. Tappert, O. Spormann, J. Fricke, *J. Appl. Cryst.* **24** (1991) 777.
169. T. Woignier, J. Phalippou, H. Hdach, G. W. Scherer in [19, d], p. 1087.
170. K. Tajiri, K. Igarashi, T. Nishio in [10], p. 83.
171. N. Hüsing, F. Schwertfeger, W. Tappert, U. Schubert in [10], p. 37.
172. T. Woignier, J. Phalippou in M. D. Baro, N. Clavaguera (eds.): *Proc. 1st Int. Workshop on Non-Crystalline Solids*, San Feliu de Guixols, World Scientific, Singapore 1986, p. 415.
173. M. Prassas, L. L. Hench in L. L. Hench, D. R. Ulrich (eds.): *Ultrastructure Processing of Ceramics, Glasses and Composites*, J. Wiley & Sons, New York 1984, p. 100.
174. T. Woignier, J. Phalippou, J. F. Quinson, M. Pauthe, F. Laveissiere in [9], p. 25.
175. P. H. Tewari, A. J. Hunt, K. D. Lofftus, *Mater. Lett.* **3** (1985) 363. Links
176. U.S. Department of Energy, US 4 610 863, 1986 (P. H. Tewari, A. J. Hunt).
177. P. H. Tewari, A. J. Hunt, K. D. Lofftus in [7], p. 31.
178. M. J. van Bommel, A. B. de Haan, *J. Mat. Sci.* **29** (1994) 943; in [10], p. 78.
179. C. J. Brodsky, E. I. Ko in [10], p. 88.
180. E. Degn. Egeberg, J. Engell in [8], p. 23.
181. D. Klvana, J. Chaouki, M. Repellin-Lacroix, G. M. Pajonk in [8], p. 29.
182. G. M. Pajonk in [8], p. 13.
183. D. M. Smith, R. Deshpande, C. J. Brinker in [19, e], p. 567.
184. Univ. of New Mexico, WO 94/25149, 1994 (R. Deshpande, D. M. Smith, C. J. Brinker).
185. D. M. Smith, D. Stein, J. M. Anderson, W. Ackermann in [10], p. 104.
186. Sinvent, WO 92/20623, 1992 (M. A. Einarsrud, S. Hæreid).
187. S. Hæreid, M. Dahle, S. Lima, M. A. Einarsrud, in [10], p. 96.
188. F. Schwertfeger, *Future Special Science — das Hoechst Magazin* **2** (1997) 28.
189. R. Saliger, U. Fischer, C. Herta, J. Fricke in [11], p. 81.
190. R. Brandt, J. Fricke in [12], p. 131.
191. R. Sempere, D. Bourret, A. Sivade, R. Etienne, J. Bouaziz, *J. Non-Cryst. Solids* **148** (1992) 499.
192. L. W. Hrubesh, J. F. Poco, *Mat. Res. Soc. Symp. Proc.* **371** (1995) 195.
193. D. M. Smith, J. Anderson, C. C. Cho, B. E. Gnade, *Mat. Res. Soc. Symp. Proc.* **371** (1995) 261. S. S. Prakash, C. J. Brinker, A. J. Hurd, *J. Non-Cryst. Solids* **190** (1995) 264; Links *Nature* **374** (1995) 439. Links
194. M. Mielke, private communication, cited in [4] .

195. A. J. Hunt, W. Cao in [19, f], p. 451.
196. W. Cao, X.-Y. Song, A. J. Hunt, *Mat. Res. Soc. Symp. Proc.* **349** (1994) 87.
197. X.-Y. Song, W. Cao, A. J. Hunt, *Mat. Res. Soc. Symp. Proc.* **394** (1994) 269.
198. X.-Y. Song, W. Cao, M. R. Ayers, A. J. Hunt, *J. Mater. Res.* **10** (1995) 1995.
199. A. J. Hunt, M. R. Ayers, W. Cao, *J. Non-Cryst. Solids* **185** (1995) 227. Links
200. D. Lee, P. C. Stevens, S. Q. Zeng, A. J. Hunt in [10], p. 285.
201. J. Kuhn, T. Gleissner, M. C. Arduini-Schuster, S. Korder, J. Fricke in [10], p. 291.
202. P. Wang, J. Kuhn, X. Lu in [10], p. 296.
203. F. Schwertfeger, U. Schubert, *Chem. Mater.* **7** (1995) 1909. Links J. Kuhn *et al.*, *Thermal Conduct.* **22** (1994) 589.
204. W. Cao, A. J. Hunt, *Appl. Phys. Lett.* **64** (1994) 2376. Links
205. M. R. Ayers, X.-Y. Song, A. J. Hunt, *J. Mat. Sci.* **31** (1996) 6251.
206. S. Reul, W. Richter, D. Haarer in [9], p. 149.
207. R. W. Pekala, S. T. Mayer, J. L. Kaschmitter, F. M. Kong in [18], p. 369.
208. S. T. Mayer, R. W. Pekala, J. L. Kaschmitter, *J. Electrochem. Soc.* **140** (1993) 446. Links
209. S. T. Mayer, J. L. Kaschmitter, R. W. Pekala in B. M. Barnett *et al.*: "New Sealed Rechargeable Batteries and Supercapacitators," *ECS Symp. Proc.* **93** (1993) no. 23, 38.
210. R. W. Pekala, S. T. Mayer, J. F. Poco, J. L. Kaschmitter, *Mat. Res. Soc. Symp. Proc.* **349** (1994) 79.
211. D. M. Smith, D.-W. Hua, W. L. Earl, *MRS-Bull.* **14** (1994) no. 4, 44.
212. S. Brunauer, P. H. Emmett, E. Teller, *J. Am. Chem. Soc.* **60** (1938) 309. Links
213. S. J. Gregg, S. W. Sing, *Adsorption, Surface Area and Porosity*, Academic Press, New York 1982.
214. G. W. Scherer, D. M. Smith, D. Stein in [10], p. 309.
215. T. Woignier, J. Phalippou, *J. Non-Cryst. Solids* **93** (1987) 17. Links
216. E. W. Washburn, *Proc. Nat. Acad. Sci.* **7** (1921) 115.
217. G. A. Nicolaon, S. J. Teichner, *Bull. Soc. Chim.* **9** (1968) 3555.
218. J. Fricke, G. Reichenauer, *J. Non-Cryst. Solids* **95 & 96** (1987) 1135.
219. R. A. Assink, B. D. Kay, *J. Non-Cryst. Solids* **99** (1988) 359. Links
220. T. Ho, J. Fraissard, *J. Chem. Phys.* **76** (1982) 5225.
221. F. Devreux, J. P. Boilot, F. Chaput, B. Savopal, *Phys. Rev. Lett.* **65** (1990) 614. Links
222. D. W. Schaefer in [8], p. 121.
223. A. Borne *et al.*, *J. Non-Cryst. Solids* **188** (1995) 235. Links
224. P. Wang *et al.* in [9], p. 141.
225. P. H. Tewari, A. J. Hunt, K. D. Lofftus, J. G. Lieber in [19, b], p. 195.
226. G. Poeltz, R. Riethmüller, *Nucl. Instr. Meth.* **195** (1982) 491.
227. S. Henning, L. Svensson, *Phys. Scr.* **23** (1981) 697.
228. R. E. Russo, A. J. Hunt, *J. Non-Cryst. Solids* **86** (1986) 219. Links
229. A. Emmerling *et al.*, *J. Non-Cryst. Solids* **185** (1995) 240. Links
230. A. Emmerling, P. Wang, G. Popp, A. Beck, J. Fricke, *J. Phys. (Paris) IV* **3** (1993) C8–360.
231. W. Cao, A. J. Hunt, *J. Non-Cryst. Solids* **176** (1994) 18. Links
232. C. A. Morris, D. R. Rolison, K. E. Swider-Lyons, E. J. Osburn-Atkinson, C. I. Merzbacher in [12], p. 29.
233. A. J. Hunt, K. D. Lofftus, *Mat. Res. Soc. Symp. Proc.* **121** (1988) 3146.
234. T. Woignier, J. Phalippou, R. Vacher, *J. Mater. Res.* **4** (1989) 688. Links
235. S. J. Kramer, F. Rubio-Alonso, J. D. Mackenzie in [19, g], p. 295.
236. M. Gronauer, J. Fricke, *Acustica* **59** (1986) 177. Links
237. M. Gronauer, A. Kadur, J. Fricke in [7], p. 167.
238. J. Gross, R. Goswin, R. Gerlach, J. Fricke in [8], p. 185.
239. J. Fricke, *Spektrum der Wissenschaft* **7** (1988) 60, **11** (1992) 14; *Sci. Am.* **256** (1988) no. 5, 92.
240. P. Scheuerpflug, R. Caps, D. Büttner, J. Fricke, *Int. J. Heat Mass Transfer* **28** (1985) 2299.
241. X. Lu *et al.* in [9], p. 207.
242. J. Kuhn, F. Schwertfeger, M. C. Arduini-Schuster, J. Fricke, U. Schubert in [10], p. 184.
243. L. W. Hrubesh in [11], p. 335.
244. G. Poeltz in [7], p. 176.
245. V. Wittner in [9], p. 233.
246. J. Fricke, *Physik in unserer Zeit* **20** (1989) 189.
247. J. Fricke *et al.*, *Thermal Cond.* **21** (1990) 235.
248. J. Fricke, X. Lu, P. Wang, D. Büttner, U. Heinemann, *Int. J. Heat Mass. Transfer* **35** (1992) 2305.
249. K. I. Jensen in [9], p. 237.
250. J. Fricke, *J. Sol. Energy Mater.* **16** (1987) 267.
251. L. F. Jesch: *Transparent Insulation Materials for Passive Solar Energy Utilization*, Franklin Comp. Cons. Ltd., Birmingham 1987.
252. M. Rommel, V. Wittwer, A. Goetzberger, *Adv. Sol. Energy Technol.* **4** (1988) 1553.

253. A. J. Hunt, K. D. Lofftus, *Adv. Sol. Energy Technol.* **4** (1988) 3146.
254. BASF, EP 340 707, 1989 (B. Wolff, G. Seybold, F. E. Krueckau).
255. W. C. Conner, G. M. Pajonk, S. J. Teichner, *Adv. Catal.* **34** (1986) 1.
256. A. J. Fanelli, J. V. Burlew, G. B. Marsh, *J. Catal.* **116** (1989) 318. Links
257. R. J. Willey, J. Olmstead, V. Djuhadi, S. J. Teichner, *Mat. Res. Soc. Symp. Proc.* **111** (1988) 359.
258. Y. Mizushima, M. Hori, *Appl. Catal.* **88** (1992) 137.
259. D. W. Cooper, *Part. Sci. Technol.* **7** (1989) 371.
260. C. Liu, S. Komarneni, *Mat. Res. Soc. Symp. Proc.* **371** (1995) 217.
261. S. Komarneni, R. Roy, U. Selvaraj, P. B. Malla, E. Breval, *J. Mater. Res.* **8** (1993) 3163. Links
262. Y. A. Attia, M. S. Ahmed, M. Zhu in [18], p. 311.
263. B. C. Dave, B. Dunn, J. S. Valentine, J. I. Zink, *Mat. Res. Soc. Symp. Proc.* **431** (1996) 285.
264. S. T. Teichner, M. Khalfallah, D. Bianchi, J.-L. Gass in [18], p. 323.
265. T. M. Tillotson *et al.* in [11], p. 358.
266. J. C. Farmer, D. V. Fix, G. V. Mack, R. W. Pekala, J. F. Poco, *J. Electrochem. Soc.* **143** (1996) 159. Links
267. R. W. Pekala *et al.* in [11], p. 74.
268. B. Pommier, S. J. Teichner, P. Lejay, A. Sulpice, R. Tournier in [8], p. 41.
269. K. E. Swider, C. I. Merzbacher, P. L. Hagans, D. R. Rolison, *Chem. Mater.* **9** (1997) 1248. Links
270. P. Löbmann, W. Glaubitt, J. Gross, J. Fricke in [10], p. 59.
271. J. Gross, J. Fricke, L. W. Hrubesh, *J. Acoust. Soc. Am.* **91** (1992) 2004.
272. V. Gibiat, O. Lefeuvre, T. Woignier, J. Pelous, J. Phalippou in [10], p. 244.
273. A. Zimmermann, J. Gross, J. Fricke in [10], p. 238.
274. M. Schmidt, F. Schwertfeger in [11], p. 364.
275. J. Phalippou, T. Woignier, J. Zarzycki in L. L. Hench, D. R. Ulrich (eds.): *Ultrastructure Processing of Advanced Ceramics, Glasses and Composites*, J. Wiley & Sons, New York 1984, p. 70.
276. M. Prassas, J. Phalippou, J. Zarzycki in L. L. Hench, D. R. Ulrich (eds.): *Science of Ceramic Chemical Processing*, J. Wiley & Sons, New York 1986, p. 156.
277. T. Woignier, J. Phalippou, M. Prassas, *J. Mater. Sci.* **25** (1990) 3118. Links
278. C. A. M. Mulder, J. G. van Lierop, G. Frens, *J. Non-Cryst. Solids* **82** (1986) 92. Links
279. E. Papanikolau *et al.*, *J. Non-Cryst. Solids* **100** (1988) 247. J. Iura, T. Kawaguchi in [8], p. 53. Links
280. M. N. Rahaman, L. C. de Jonghe, S. L. Shinde, P. H. Tewari, *J. Am. Ceram. Soc.* **71** (1988) C338. Links
281. A. Emmerling *et al.*, *J. Non-Cryst. Solids* **125** (1990) 230. Links
282. R. Sempere, D. Bourret, J. Bouaziz, A. Sivade in [8], p. 227.
283. B. Dunn, J. I. Zink, *J. Mater. Chem.* **1** (1991) 903. Links
284. T. M. Tillotson, W. E. Sunderland, I. M. Thomas, L. W. Hrubesh, *J. Sol – Gel Sci. Technol.* **1** (1994) 241.
285. A. Charlton, I. T. McKinnie, M. A. Meneses-Nava, T. A. King, *J. Mod. Opt.* **39** (1992) 1717.
286. C. S. Ashley *et al.* in [22], p. 989.
287. L. W. Hrubesh, J. F. Poco, *J. Non-Cryst. Solids* **188** (1995) 46. Links
288. W. Ebeling, D. A. Reierson, R. J. Pence, M. S. Viray, *Pestic. Biochem. Physiol.* **5** (1975) 81. Links
289. P. K. Sen-Sarma, S. C. Mishra, *Indian J. Entomol.* **38** (1976) 77.
290. C. H. Vrba, H. P. Arai, M. Nosal, *Can. J. Zool.* **61** (1983) 1481. Links
291. S. R. Loschiavo, *J. Econ. Entomol.* **81** (1987) 1231; **81** (1988) 1237.
292. P. Tsou, D. E. Brownlee, A. L. Albee, *LPSC* **19** (1988) 1205.
293. P. Tsou in [10], p. 415.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Centrifuges, Filtering

Standard Article

Urs A. Peuker[1]

[1]Clausthal University of Technology, Clausthal-Zellerfeld, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2007 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.c05_c01.pub2
Article Online Posting Date: January 15, 2007

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Fundamentals
2.1. Description of the Pore Liquid
2.2. Cake Formation
2.3. Washing of the Centrifuge Cake
2.4 Dewatering of the Centrifuge Cake
2.4.1. Equilibrium of Centrifugal Dewatering
2.4.2. Kinetics of Dewatering
3. Discontinuous Filtering Centrifuges
3.1. Cycles of the Discontinuous Centrifugation Process
3.2. Filling Strategies of the Centrifuge Drum
3.3. Peeler/Scraper Centrifuge
3.3.1. Horizontal Construction
3.3.2. Vertical Construction: Three Column Centrifuge
3.3.3. Siphon Centrifuge
3.4. Inverting Filter Centrifuge
3.5. Centrifugal Dryer
4. Filtering Centrifuges with Continuous Feed
4.1. Conical Screen Centrifuge
4.2. Vibrating Conical Screen Centrifuge
4.3. Tumbler Centrifuge
4.4. Screen Scroll Centrifuge
4.5. Pusher Centrifuge
4.5.1. Cake conveyance – Transport model
4.5.2. Washing on the Pusher Centrifuge
4.5.3. Feeding – Danger of Basket Flooding
4.5.4. Multistage Pusher Centrifuges
4.5.5. Special Pusher Centrifuges
4.6. Decanting Centrifuge/Solid Bowl Centrifuge
4.6.1. Application for Filtration Dewatering
4.6.2. Dewatering Model
4.7. Screen-Bowl Centrifuge
5. Hybrid Processes
5.1. Hyperbaric Centrifugation/Jet-Stream Centrifugation
5.2. Steam-Enhanced Centrifugation/Steam-Pressure Centrifugation
6. Laboratory-Scale Testing
6.1. Laboratory Centrifuges

7. Applications
8. References

[Top of Page]

1. Introduction

Centrifuges are machines which are primarily used in solid–liquid separation. The driving force is the centrifugal field, which acts differently on the particles and the mother liquid. Filtering centrifuges compete with filters (⟶ Filtration), press filters, and sedimenting centrifuges (⟶ Centrifuges, Sedimenting).

The field of application for filtering centrifuges is quite wide: minerals, fine chemicals, pharmaceutical products, food products, and bio suspensions. Therefore, different concepts for construction have been developed. Furthermore, most centrifuges are adapted in detail engineering to their application. Filtering centrifuges can be operated in continuous or discontinuous mode. The capacity can range from several kilograms up to more than 100 t per hour.

All filtering centrifuges have in common the dewatering process. First, a filter cake is formed on a permeable support, second, the liquid drains in the direction of the centrifugal field through the filter cake; third, desaturation begins, the liquid runs off the pores, and air penetrates the pores of the cake. The process comes to an end when the driving force of the centrifugal field is in equilibrium with the retaining forces in the pores, which primarily are due to wetting effects (Fig. 1).

Figure 1. Process phases during centrifugal filtration: sedimentation, drainage, and desaturation

a) Liquid surface; b) Clear liquid; c) Sedimenting suspension; d) Filtrate; e) Clear liquid; f) Centrifuge cake; g) Desaturated cake; h) Wet cake

The main component of a filtering centrifuge is the perforated drum or basket. Its interior is covered with a sieve or a filter medium. The rotation of the basket usually is powered by an electric drive. The performance of a centrifuge can be estimated by using the centrifugation number *C*, which is the ratio of the centrifugal acceleration to gravity. It depends on the angular velocity ω and the basket diameter *R*.

$$C = \frac{R\,\omega^2}{g} \tag{1}$$

Technical filtering centrifuges operate with a centrifugal number from about 100 up to 2000 or even higher (Fig. 2). The maximum rotary speed is limited by the mechanical stress and the durability of the steel used for construction.

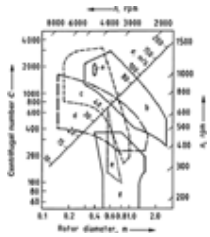

Figure 2. Operational ranges for filtering centrifuges

a) Inverting filter; b) Peeler centrifuge; c) Screen scroll; d) Pusher centrifuge; e) Tumbler centrifuge; f) Pendulum centrifuge

The vertical lines perpendicular to the rotor-diameter axis and the rotational speed (*n*) lines from top right to bottom left intersect at the operating point. The ordinate of this intersection gives the centrifugal number *C*. Secondary parameter lines from top left to bottom right give the circumferential velocity *v*.

Table 1 gives an overview on the construction size and performance of the common types of filtering centrifuges. The solids throughput depends mainly on the suspension and particle properties, which must be estimated with laboratory-scale tests.

Table 1. Types of centrifuges: constructive and operation parameters

Centrifuge type	Rotor diameter, mm [1], [2]	Throughput, t/h [2], [3]	Maximum centrifugal number	Minimum particle size, µm [1], [3]	Minimum solids concentration, % [1], [3]
Peeler centrifuge	250–2000	0.1–15	1500 (2000)	5–10	10
Inverting filter centrifuge	400–1300		700–1100	2	5
Centrifugal dryer	400–1300		600–700	2	5
Conical screen centrifuge	500–800		1250–2400	50–80	15
Vibrating conical screen centrifuge		300	30–120	300	50

Tumbler centrifuge		200	50–300	200	50
Screen scroll centrifuge	200–1000	0.1–100	500–1500	80	20
Pusher Centrifuge	200–1560	2–100	300–1200	50	20
Decanting centrifuge			> 4000	70	4–5
Screen-bowl centrifuge	360–1000	0.2–30	1800–3200	20	4–5

[Top of Page]

2. Fundamentals

2.1. Description of the Pore Liquid

The pore liquid in a filter or centrifuge cake consists of differently bound liquid (Fig. 3). The main fraction is the liquid in the large interparticle pores (a), which are easily accessible. After drainage of the pore liquid, liquid bridges (b) remain within the porous system at the contact faces of the particles. They are bound by strong surface forces, as is liquid in the surface roughness of the particles (c). The inner porosity of the solid also contains liquid (d). Any particle surface is covered by a liquid surface film as well as with adsorbed liquid (e). Centrifugal dewatering is able to remove the liquid from the large pores. With large particle sizes some of the liquid bridges also can be removed by the centrifugal force. All other liquid cannot be removed by mechanical means in a centrifugal field or by filtration.

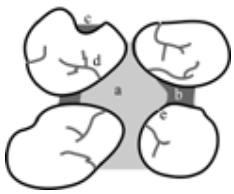

Figure 3. Pore liquid binding mechanisms in the porous matrix

a) Pore liquid; b) Liquid bridges; c) Liquid in surface roughness; d) Liquid in inner porosity; e) Adsorbed liquid film

The centrifuge cake is described by its geometry in the basket and its inner structure. The porosity ε is the ratio of void volume V_v in the cake to the entire cake volume V_c.

$$\varepsilon = \frac{V_v}{V_c} \tag{2}$$

Dewatering of the porous cake is quantified with different dimensionless numbers: the saturation S, which is the ratio of liquid volume to pore volume, the liquid load X, which is the ratio of liquid mass to solids mass, and the residual moisture R_M, which is the ratio of liquid mass to mass of wet cake.

$$S = \frac{V_L}{V_v} = \frac{V_L}{V_S \cdot \frac{\varepsilon}{1-\varepsilon}} \quad X = \frac{m_L}{m_S} \quad R_M = \frac{m_L}{m_L + m_S} \tag{3}$$

Due to its dependence on porosity, saturation is a more academic unit. Residual moisture and liquid load are derived from practical application.

2.2. Cake Formation

Cake formation in the drum must take account of the curvature of the filtering surface. The cake is built up as an annulus (Fig. 4). Thus, the area through which the liquid drains is a function of cake height. This must be considered when calculating the pressure drop.

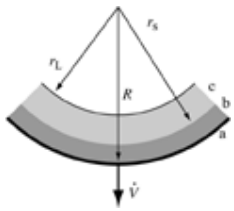

Figure 4. Cake geometry

a) Filter medium; b) Cake; c) Clear liquid pool

The volume flux $\dot{V}$ can be written using the Darcy equation:

$$\dot{V} = -\frac{2\,\pi\, r L_{dr}}{\eta_L\, r_c} \cdot \frac{dp}{dr} \tag{4}$$

where r is the radius, L_{dr} the length of the drum, p the pressure, η_L the dynamic liquid viscosity, and r_c the specific cake

resistance.

Integration within the boundaries of drum radius R and radius of the cake surface r_S leads to the pressure drop as a function of the cake height (Eq. 5).

$$\Delta p = -\frac{\dot{V}\eta_L r_c}{\pi r L_{dr}} \cdot \ln\frac{R}{r_s} \tag{5}$$

This pressure drop of the volume flux is balanced by the driving centrifugal pressure of the liquid in the basket (Eq. 6). The liquid in the pores of the cake and the clear liquid in the pool above the cake surface r_L must be considered:

$$\Delta p = \frac{1}{2} \cdot \rho_L \omega^2 (R^2 - r_L{}^2) \tag{6}$$

where ρ_L is the liquid density.

The force balance when the pressure drop is equal to the driving centrifugal pressure allows the volume flux of filtrate draining off the cake to be determined (Eq. 7).

$$\dot{V} = \frac{\pi L_{dr} \rho_L \omega^2 (R^2 - r_L{}^2)}{\eta_L r_C \ln\frac{R}{r_S}} \tag{7}$$

In the case of small cake heights the equation can be simplified by replacing the logarithm (Eq. 7) by a linear term (Eq. 8).

$$\dot{V} = \frac{\pi L_{dr} \rho_L \omega^2 (R^2 - r_L{}^2)}{\eta_L r_C (R - r_S)} \tag{8}$$

The volume flux can be written with separated groups of parameters (Eq. 9): filtering surface, centrifugal acceleration, product parameter, and cake geometry.

$$\dot{V} = A_c C g \cdot \frac{\rho_L}{2 \eta_L r_C} \cdot \frac{(R^2 - r_L{}^2)}{R^2 \ln\frac{R}{r_S}} \tag{9}$$

The volume flux can be used to calculate the draining time in a centrifuge (Eq. 10). The drum usually is filled with a flow rate higher than the draining rate. Cake formation is an instantaneous process and the liquid remains above the cake. Subsequent to draining, a second filling step, a washing step, or desaturation occurs. With the focus on filling and washing it is necessary to predict the right moment when the liquid level reaches the cake surface:

$$\Delta t = \frac{\eta_L r_C \ln\frac{R}{r_S}}{\rho_L \omega^2} \cdot \ln\left(\frac{R^2 - r_{L,t=0}{}^2}{R^2 - r_S{}^2}\right) \tag{10}$$

The estimation of the filling time at the optimum feed flux, which is a function of the cake height in the basket, requires an extended mass balance [1].

2.3. Washing of the Centrifuge Cake

Cake washing is an essential process step to increase product quality. The mechanism and process design differ from continuous to discontinuous centrifuges. Continuous centrifuges are characterized by low specific washing liquid quantities from 20 to 200 kg/t_{solids}. Washing mainly is a film washing process. Film washing has its highest efficiency when the wash liquid is injected directly after cake formation, when the centrifuge cake has its highest saturation. The more liquid has drained from the pores, the more closed liquid compartments are created within the porous system. These are protected by the surface tension and therefore they are not directly accessible for the wash liquid. The wash liquid itself is sprayed onto the cake surface through pressure nozzles. The residence time of the cake in the centrifuge must be high enough to ensure that both the pore and the wash liquid are able to drain out of the cake. Usually the wash liquid prolongs the film flow phase of dewatering [4].

Discontinuous centrifuges use a washing mechanism which is similar to that used in filtration. The wash liquid flows in plug flow through the entirely saturated filter cake. The wash ratio, defined as was liquid per unit pore volume, in most cases is greater than one.

The washing step can be operated at lower rotation speed than dewatering. The liquid is softly laminated above the cake surface, which must be disturbed and resuspended as little as possible.

The wash liquid then drains through the cake. The flux can be calculated using the same equation as for the cake formation process (Eq. 7).

2.4. Dewatering of the Centrifuge Cake

The dewatering process can be split up into different subprocesses. Feeding of the suspension into the drum is directly coupled to cake formation. Cake formation by sedimentation is characteristic of centrifugal processes in which the solid density is higher than the liquid density. Formation of the cake is as instantaneous effect.

Subsequent to cake formation the liquid is forced by its own weight to drain through the porous sediment. Desaturation occurs when the surface forces can be overcome by the mass forces. Finally the dewatering process reaches an

equilibrium state.

2.4.1. *Equilibrium of Centrifugal Dewatering*

The driving force of the centrifugation process is the centrifugal pressure p_c (Eq. 11), which is calculated from the liquid density and the height of the liquid in the drum *H*:

$$p_c = \rho_L g C H \tag{11}$$

Using the approximation that the cake height is small compared to the drum diameter a simple force balance can be applied (Fig. 5). The filter cake is replaced by a capillary with a representative diameter. This diameter is called the hydraulic diameter (Eq. 12).

$$d_h = \sqrt{\frac{32\,\varepsilon}{r_c}} \tag{12}$$

where r_c is the specific filter cake resistance, and ε porosity.

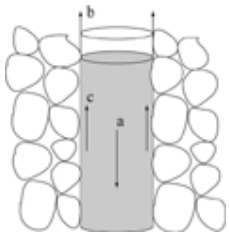

Figure 5. Force balance in a characteristic capillary

a) Centrifugal force; b) Surface force; c) Pressure drop (shear force)

The driving force is the centrifugal force and the retaining force is the surface force:

$$\pi \cdot \frac{d_h}{4} \cdot \rho_L\, g C H - \pi\, d_h\, \sigma \cos \delta = 0 \tag{13}$$

Transformation of Equation (13) leads to the definition of the Bond number Bd_1 (Eq. 14):

$$Bd_1 = \frac{\rho_L\, g C d_h H}{\sigma \cos \delta} \tag{14}$$

The Bond number Bd_1 also can be seen as a reciprocal relative liquid height in the capillary. The Bond diagram (Fig. 6) shows the equilibrium of dewatering: Saturation is plotted versus Bd_1. Information on the minimal saturation for a given set of operation parameters can be drawn from the Bond diagram.

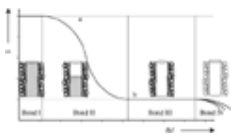

Figure 6. Equilibrium of dewatering, characteristic Bond-diagram

a) Saturation *S* = 1 — no dewatering; b) Bond plateau

The Bond diagram (Fig. 6) contains four characteristic areas:

- Bond I — no desaturation occurs, saturation is equal to one, and the surface force equals the centrifugal force.
- Bond II — desaturation begins, and with increasing *Bd* saturation decreases until all liquid is removed from the large interparticular pores.
- Bond III — the so-called Bond plateau; saturation remains constant, and to drain the liquid bridges the centrifugal force must be increased by at least one order of magnitude.
- Bond IV — this area is usually out of range for technical dewatering processes; the liquid bridges start to drain.

Filtering centrifuges are operated in Bond I to III, and continuous filtering centrifuges mainly in Bond III. Discontinuous filtering centrifuges operated in Bond II create a heterogeneous distribution of liquid within the cake. The cake near the filter cloth is saturated, and the cake near the surface is desaturated.

The description of the area Bond IV also uses a second Bond number (Eq. 15), which takes into account that the cake height is no longer an influencing parameter. The dimensionless number Bd_2 is calculated with the square of the hydraulic diameter, which is a characteristic length for the liquid bridges within the pores.

$$Bd_2 = \frac{\rho_L\, g C {d_h}^2}{\sigma \cos \delta} \tag{15}$$

The Bond diagram can also be plotted versus Bd_2, and different experimental results the form a master curve in the area Bond IV but spread in the area Bond II, where the cake height has its main influence on desaturation.

2.4.2. *Kinetics of Dewatering*

Estimation of dewatering kinetics is relevant to continuous centrifuges, which have a limited residence time from about 2 to 30 s. With the help of the capillary model (Fig. 5) the drainage of the liquid can be calculated taking into account the friction force of a laminar flow within the capillary. The force balance (Eq. 16) then includes friction, surface, and centrifugal forces:

$$-8\pi\ \eta\, x \cdot \frac{\mathrm{d}x}{\mathrm{d}t} - \pi\, d_\mathrm{h}\, \sigma\, \cos\delta + \pi \frac{{d_\mathrm{h}}^2}{4} \cdot \rho_\mathrm{L}\, g C x = 0 \tag{16}$$

The solution of the force balance leads to an implicit equation of the dewatering time and the time dependent position *x* of the liquid surface within the capillary (Eq. 17). The left side of the equation, which contains the time Δt, can be merged into the kinetic parameter. This kinetic parameter λ can be seen as a reciprocal time.

$$\frac{1}{8\lambda} = \frac{\Delta t \rho_\mathrm{L} g C d_\mathrm{h}^2}{32\eta H} = \left(1 - \frac{x}{H}\right) - \frac{4\sigma\cos\delta}{\rho_\mathrm{L} g C H d_\mathrm{h}^2} \cdot \ln\left[\frac{\frac{x}{H} - \frac{4\sigma\cos\delta}{\rho_\mathrm{L} g C H d_\mathrm{h}^2}}{1 - \frac{4\sigma\cos\delta}{\rho_\mathrm{L} g C H d_\mathrm{h}^2}}\right] \tag{17}$$

In the force balance of the plug flow (Eq. 16) the surface forces also can be expressed by the measured capillary pressure of the centrifuge cake (Eq. 18). This approach decouples the hydrodynamic properties of the hydraulic diameter from its effect on the surface force. The decoupling becomes more and more necessary the smaller the hydraulic diameter gets, because for fine particles the surface force calculated using the hydraulic diameter increases more than in the real cake.

$$\frac{1}{8 \cdot \lambda} = \frac{\Delta t \cdot \rho_\mathrm{L} \cdot g \cdot C \cdot d_\mathrm{h}^2}{32 \cdot \eta \cdot H} = \left(1 - \frac{x}{H}\right) - \frac{p_\mathrm{ke}}{\rho_\mathrm{L} \cdot g \cdot C \cdot H} \cdot \ln\left[\frac{\frac{x}{H} - \frac{p_\mathrm{ke}}{\rho_\mathrm{L} \cdot g \cdot C \cdot H}}{1 - \frac{p_\mathrm{ke}}{\rho_\mathrm{L} \cdot g \cdot C \cdot H}}\right] \tag{18}$$

The calculation of the plug flow determines the moment when the single-phase flow of the pore liquid has reached the filter media. From then on the remaining pore liquid drains as surface or film flow in the capillaries of the cake.

The regime of film flow can be described by applying the Nusselt film theory of a liquid film draining down a vertical plate. Finally, the saturation within the pores of the cake can be expressed with the same kinetic parameter λ introduced for plug flow (Eq. 17). The film saturation can be written as a power law with two optional parameters *a* and *b* (Eq. 19). The rigorous solution for a smooth and even surface quantifies *a* to 4/3 and *b* to 1/2. Usually *b* is a product parameter and it depends on the particle size distribution and the roughness of the particles. Therefore the numerical value of *b* after fitting to experimental data can vary in the range from 0.27 to 0.6.

$$S_\mathrm{F} = a \cdot \lambda^b \tag{19}$$

The time-dependent saturation of the centrifuge is calculated by a combination of both approaches. The overall saturation consists of three contributions: the mechanically nonremovable liquid, the film flow, and the plug flow (Eq. 20, Fig. 7). The nonremovable liquid consists of the liquid bridges, the adsorbed liquid, and the liquid located in inner porosity. The nonremovable saturation S_non amounts to 0.06 to 0.09 for smooth spheres. A typical value of S_non for mineral products is about 0.15.

$$S(t) = S_\mathrm{non} + \left(1 - S_\mathrm{non}\right) \cdot \left[\frac{x(t)}{H} + \left(1 - \frac{x(t)}{H}\right) \cdot S_\mathrm{F}\right] \tag{20}$$

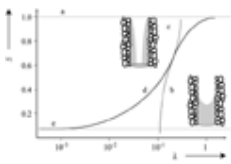

Figure 7. Characteristic time-dependent desaturation; superposition of film and plug flow

a) Maximum saturation *S* = 1; b) Plug flow *S*(*t*); c) Film flow *S*(*t*); d) Superposition; e) Equilibrium saturation

An alternative approach to calculate the time-dependent dewatering of the cake uses the pore size distribution [5], [6], which is given by a capillary pressure curve. In all classes of capillaries the plug flow is calculated, without any contribution of the film flow. But the pore size distribution is not given in most practical problems, and therefore it is difficult to apply this model.

[Top of Page]

3. Discontinuous Filtering Centrifuges

The process integration of discontinuous centrifuges aims at processes which are operated batchwise. Semicontinuous operation is possible when several discontinuous centrifuges are combined. The maximum throughput of discontinuous centrifuges is 5–10 t_solids per hour. The maximal basket diameter of discontinuous centrifuges is about 2000 mm. The common size of discontinuous filtering centrifuges ranges from 600 to 1500 mm. The maximum centrifugal number is 2000, and the typical operating range *C* = 600–1500.

3.1. Cycle of the Discontinuous Centrifugation Process

Discontinuous centrifuges are operated batchwise. The process steps are executed in succession (Fig. 8). First the centrifuge is accelerated, and then the suspension is fed into the basket. It depends on the filtration properties, the concentration of the suspension, and the throughput whether more than one filling step is necessary. The rotational speed during the filling process is lower than the maximum speed, because usually the drive is not designed for the maximum drive torque, which occurs during the acceleration of the suspension.

After the cake has built up, the centrifuge basket can be accelerated to operational speed. The next steps are one or more washing steps, if required. After the last washing step desaturation (plug and film flow according to Eq. 19) occurs. Discontinuous centrifuges do not have technical restrictions on the residence time. Desaturation can be continued until saturation has reached equilibrium or the Bond saturation.

It depends on the type of discontinuous centrifuge at which rotational speed the basket is unloaded. Modern peeler centrifuges remove the cake at quite high rotational speed; the inverting filter centrifuge requires a low speed.

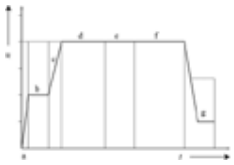

Figure 8. Cycle of discontinuous centrifugation

a) Acceleration of the basket; b) Feeding speed (≤ maximum speed); c) Acceleration to maximum speed; d) First washing step; e) Second washing step; f) Desaturation; g) Discharge at lower speed

3.2. Filling Strategies of the Centrifuge Drum

The cake formation process during filling depends on the ratio f_r of liquid fed to the basket and liquid consumed by the growing cake, either by being incorporated in the porosity or by passing through as filtrate flux. The filtration ratio can be approximated for suspensions of lower concentration by the ratio of liquid content of the feed suspension to the filtrate flux (Eq. 21). Here the liquid which fills the cake porosity is neglected.

$$fr = \frac{c_V \cdot \dot{V}_{susp}}{\dot{V}_L} \tag{21}$$

The right filling strategy must be developed to meet the requirements of the centrifuge. If the feed flux is too small, instant filtration occurs when the suspension hits the filter medium or the surface of the filter cake. In this case there is no smoothening effect of a liquid pool above the growing cake. The cake is formed where the suspension hits the surface and becomes more and more uneven, which leads to an imbalance in the basket rotation. The uneven cake height (Fig. 9) negatively affects the washing step, because the washing liquid drains through the thinner areas of the cake. The integral degree of impurity increases.

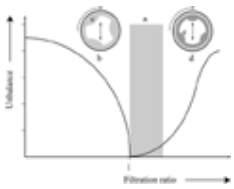

Figure 9. Regimes of cake formation as a function of feed flux

a) Optimal range of operation with minimal liquid pool; b) Low feed flux and instantaneous cake formation; c) Imbalance due to irregular cake formation; d) High feed flux and increasing pool depth; e) Imbalance due to liquid waves at the pool surface

High suspension feed rates lead to a high liquid pool above the cake, which increases the sedimentation distance of the particles. During sedimentation segregation can occur. Due to this segregation a heterogeneous cake with a higher resistance and a skin layer is formed, and this also has negative effects on the residence time and the desaturation properties. The skin layer increases the capillary entry pressure of the cake. Fine products showing this tight skin may even not be able to be desaturated in the centrifuge.

The optimum feed rate of a discontinuous basket centrifuge must be chosen between these two boundaries. In conventional processes the feed rate is constant over the filling interval, but the cake permeability and hence the filtrate flux decreases (Eq. 7). Therefore, the higher the cake grows, the deeper the overlying liquid pool becomes. Modern approaches try to set up a control strategy based on the filtration and desaturation models [7] to regulate the feed flux.

3.3. Peeler/Scraper Centrifuge

The peeler centrifuge is the most widespread construction of discontinuous centrifuges. The name derives from the strategy to remove the solids from the basket. A scraper is used which immerges into the cake while the basket still is in rotation. This automatic discharge replaces the manual labor of the operator which formerly was necessary.

Cake removal with a scraper leaves a thin cake layer in the basket. This layer is characteristic for the centrifuge type. As a positive effect it prevents any particle transfer through the filter medium in the following filling step. However, the thin layer ages due to deep-bed filtration effects and therefore its resistance increases from cycle to cycle, which leads to longer and longer processing times. When the thin layer has reached a certain resistance it is manually removed.

With the focus on hygienic design and good manufacturing practice the thin layer is a severe problem. It must be removed after each centrifugation step to prevent, e.g., any cross-contamination of separate batches.

Several concepts to remove the ground layer are available. The layer can be removed with mechanical, pneumatic, or hydraulic devices. The pneumatic concept either is operated with a strong air blow from the housing through the basket or

directly with an additional nozzle holder inside. The hydraulic concept feeds a larger quantity of liquid into the basket after scraping is finished. The ground layer is resuspended and flushed out off the basket. Any further cleaning uses the CIP (cleaning in place) equipment.

3.3.1. *Horizontal Construction*

The basket is mounted on a horizontal shaft, supported by fixed bearings. At the opposite side the shaft is powered by an electric drive via a V-belt. The basket is perforated and equipped with drainage media and a filtration medium above it. The front side can be closed with a door. An alternative concept uses a traversable housing, which can be removed to provide access to the entire basket and the internals.

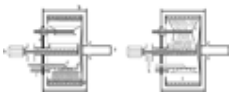

Figure 10. Peeler centrifuge, in feeding mode (left) and discharge mode (right)

a) Shaft; b) Drive of transport screw; c) Feeding pipe; d) Washing pipe; e) Centrifuge lid; f) CIP nozzle; g) Centrifuge cake in the basket; h) Housing; i) Scraper during peeling of the cake; j) Transport chute; k) Solids discharge; l) Remaining thin layer

The feed piping reaches into the basket, and a special distribution device is mounted at its end (Fig. 10). The discharge scraper and the wash pipes are also located inside the basket. The discharge scraper feeds the solid to a funnel which is connected either to a chute or a transport screw. During filling, washing, and filtration the discharge unit is turned toward the inside of the basket.

Cake washing uses wash pipes or nozzles to achieve uniform distribution of the wash liquid over the cake surface. The flux of the injected wash liquid is high enough to create a liquid pool above the cake, which slowly drains through the cake.

3.3.2. *Vertical Construction; Three-Column Centrifuge*

The vertical peeler centrifuge consists of a vertically mounted basket in which the drainage and filter media are fixed. All feed pipes pass through the lid of the centrifuge. Directly after cutting, the solids are transported either upward with a pneumatic conveyer or fall down through the backplane of the basket into a chute.

To reduce vibration the vertical centrifuge is fixed on a massive frame or plate which lies on rubber dampers. An alternative concept is the three-column centrifuge. The basket and the drive unit are mounted as a kind of pendulum on three support points. This concept reduces the vibrations transferred to the ground.

3.3.3. *Siphon Centrifuge*

The siphon centrifuge is a modified horizontal peeler centrifuge in which a siphon is connected to the filtrate piping. Due to the hydraulic contact from the pore liquid through the filter medium into the filtrate in the siphon a hydrostatic pressure is built up. This suction is an additional driving force which accelerates drainage during cake formation, washing, and the plug-flow phase of desaturation. At the moment when the surface of the plug reaches the filter medium the hydraulic contact breaks and the suction force becomes zero.

The siphon also allows backflushing of filtrate into the basket, which can be used to regenerate the ground layer. The efficiency of the siphon can directly be controlled by the position of the skimmer tube (Fig. 11). It can be of advantage to start the suction force after the feeding of the basket, because a high filtrate flux at the beginning of cake formation favors formation of an uneven surface.

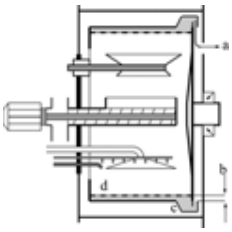

Figure 11. Principle of the siphon centrifuge

a) Skimmer tube; b) Hydraulic height; c) Siphon; d) Filter media

3.4. Inverting Filter Centrifuge

The inverting filter centrifuge is a specialized horizontal centrifuge (Fig. 12), which is used in integrated processing of high-quality products. The basket is mounted on the horizontal shaft. The basket can be closed pressure-tight with a special lid. The lid is connected by six anchors to the backplane of the basket. During the opening of the basket the connected front- and backplanes are shifted to the discharge side. The filter medium, which is fixed at the backplane and at the front edge of the basket, is inverted. After the inversion (Fig. 13) the inner side, on which the cake has been processed, has been turned outside. With a small acceleration the cake detaches from the filter medium into the housing and the discharge chute.

Due to the pressure-tight construction of the basket the inverting filter centrifuge also can be operated as hyperbaric centrifuge. The feed pipe, which is inserted from the front side though the lid, can be sealed with special membrane. The feed pipe is the only pipe which reaches into the basket; it is therefore also used to inject the washing fluid, as well as pressurized air during hyperbaric processing.

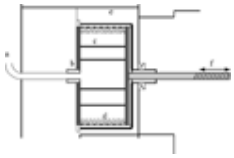

Figure 12. The inverting filter centrifuge

a) Feeding pipe; b) Membrane sealing; c) Anchors between front- and backplane; d) Cake above filter medium; e) Housing; f) Screw drive to shift the front- and backplane unit

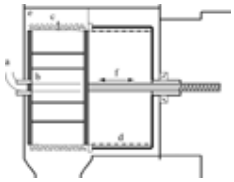

Figure 13. The inverting filter centrifuge — solids discharge

a) Feeding pipe; b) Shifted front- and backplane units; c) Removed centrifuge cake; d) Perforated basket; e) Housing and discharge chute; f) Shifted shaft

3.5. Centrifugal Dryer

The centrifugal dryer is a horizontal discontinuous centrifuge. The perforated basket is equipped with a fixed multilayer sintered metal sieve which has a minimum pore size down to 1 µm. The suspension is fed via the hollow shaft to the basket. The basket itself can be closed by a lid. The closed basket is pressure-tight up to 4.5–6 bar depending on the basket diameter. The centrifuge either is run with high rotational speed to perform centrifugal dewatering or with low speed to conduct the drying step. During the drying step air is blown through the metal sieve into the basket and the solids cake is fluidized. The product is retrieved from the basket through a connection to a pneumatic conveying system. To use this system the dry product requires excellent free-flowing properties. The centrifugal dryer is also able to perform solids washing. The washing liquid is injected through the hollow shaft.

[Top of Page]

4. Filtering Centrifuges with Continuous Feed

4.1. Conical Screen Centrifuge

Characteristic of the conical screen centrifuge is the conical basket. The suspension is fed through a nonrotating pipe to the smaller diameter of the basket. Due to the component of gravitational force parallel to the screen the solids cake is transported toward the larger basket diameter, where the discharge is located. Therefore, transport in the centrifuge depends on the slope of the conical basket and the interaction of the screen surface with the cake. At high friction between product and screen, blocking can occur.

The conical screen without any transport devices is characteristic for this centrifuge type. The lack of transporting devices allows quite high centrifugal numbers *C* of up to 3300 [1] but is also the largest disadvantage of the conical screen centrifuge, because the operator has no possibility to influence the residence time and the solids distribution in the basket.

Two constructional concepts are available for conical screen centrifuges: horizontal shaft and vertical shaft. The latter usually transport the cake upwards against gravity. The conical screen centrifuge is common in the raw sugar industry.

4.2. Vibrating Conical Screen Centrifuge

The vibrating conical screen centrifuge superimposes vibration of the basket to control and to improve the solids transport. The vibration is normal to the centrifuge axis. Therefore, the vibrating conical screen centrifuge is limited in its rotational speed because of the mechanical stress on the bearing and the shaft. Usually this centrifuge type is operated up to centrifugal numbers *C* of 120–150.

Due to this low centrifugal force only coarse products (e.g., in the mining industry) are processed with the vibrating conical screen centrifuge. The vibration can be controlled in amplitude and frequency to adjust the residence time and the product distribution in the basket.

4.3. Tumbler Centrifuge

The tumbler centrifuge superimposes a tumbling movement of the basket to improve solids transport. The tumbler centrifuge has a vertical shaft. The solids cake is transported upwards by the component of gravitational force parallel to the screen and by the revolving tumbling of the basket [8], [9]. The cake advances stepwise toward the discharge because of the quite low frequency of tumbling. The centrifugal number *C* ranges from about 150 to 300–400. The typical residence time is 5–10 s. Due to the quite high residence time and the controllable transport of the cake, washing is possible. Compared to the conical screen centrifuge the cake thickness is higher.

Tumbler centrifuges have been commonly used in the mining industry and in processing of coarse crystals.

4.4. Screen Scroll Centrifuge

The conical basket of a screen centrifuge is equipped with a conical screw to control solids transport, cake distribution, and cake thickness. This helical screw conveyor or a set of helical blades primarily transports the solids cake forward to the discharge side. The screw is operated with a differential rotation speed, which can be set up by a cyclo gear or a gearbox [3].

The screen scroll centrifuge is constructed either with vertical or horizontal shaft. The advantage of the vertical concept is

easy demounting, but the accessibility of the components during maintenance is poor. Today the horizontal concept is favored. The centrifuge either uses screens similar to the pusher centrifuge or laser-perforated plates [10]. The holes (80 µm to 1 mm) in these plates allow the processing of even finer products compared to the use of conventional screens. The void area of a typical wedge-wire screen is 4–10 % overall.

Due to the inclination α = 10–20°) of the basket [2] the solid cake slides down the screen. The screw controls this movement and with it the residence time (Eq. 22)

$$t_{res} = \frac{1}{\Delta n} \cdot \frac{L_{dr}}{H_{fl} \cos \alpha} \tag{22}$$

where H_{fl} is the flank lead of the screw, and Δn the differential speed of screw and basket. The gap between screw and basket must be quite small to prevent solids accumulation on the screen. A small gap also ensures better sealing when suspensions of low concentration are processed.

The screen scroll centrifuge has limited washing capacity. The washing liquid must be sprayed from the inside of the transport screw through holes or washing nozzles onto the solids cake.

The basket has two main functions: Near the inner diameter the basket is used as a thickening zone. Here a large quantity of liquid drains through the screen and the cake is about to be built up. Due to the ongoing concentration process the solids particles are mobile and the filtrate flux is able to transport them through the screen gap. In this part of the centrifuge the solids content in the filtrate is high.

The area close to the discharge side is used for cake dewatering, which is described by piston and film flow (Eq. 20). Due to the quite large particle diameter the efficiency of film flow is poor. The cake height is not constant within the spiral of the transport screw, but an average cake height (Eq. 23) can be calculated:

$$H = \frac{m_S t_{res}}{\rho_S (1 - \varepsilon) A_{scr}} \tag{23}$$

4.5. Pusher Centrifuge

The conventional concept of a pusher centrifuge is the horizontal machine (Fig. 14). The housing is divided into the filtrate-collecting zone and the solids discharge. The pusher centrifuge is fed from the front side. Usually the stationary tubing reaches into the drum. The suspension is fed into the drum, and in most cases an accelerating device (chamber, cone, or disk) is used. The centrifuge is fed continuously with a constant suspension flux.

The drum itself consists of a rotating perforated basket equipped with a wedge-wire screen. At the solids discharge the drum is open; at the opposite side it is connected to the hollow drive shaft. The pusher plate is located opposite to the discharge side; it acts as a backplane of the process chamber. It is mounted on its own shaft and rotates with the same speed as the basket. Additionally, the pusher plate undergoes a cyclic horizontal movement at a frequency of about 1 Hz.

The suspension is injected in the annulus which is formed by the amplitude of the pusher plate. When the pusher plate retreats, the suspension directly hits the metal sieve. This is the maximum mechanical stress for the particles [11].

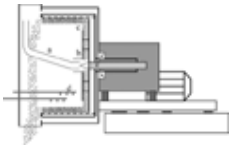

Figure 14. Single stage pusher centrifuge

a) Feeding pipe; b) Feed accelerator; c) Drainage zone; d) Washing pipe

The cyclic movement of the pusher plate initiates horizontal cake transport. The centrifugal cake slides over the metal sieves of the drum. An annulus of the cake is discharged when it is transported over the edge of the drum.

An electric motor is used to power the rotation of the drum. The cyclic movement of the pusher plate is either created by a hydraulic system or a crank lever. The hydraulic system is favored for larger machine size.

4.5.1. *Cake Conveyance — Transport Model*

The cyclic movement of the pusher plate can be divided into four phases: advancing and retreating of the pusher plate as well as front and rear dead center. For hydraulic pushing systems the advancing and retreating movement occurs with constant velocity. The position of the pusher plate influences cake formation and cake structure.

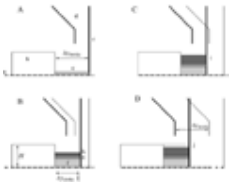

Figure 15. Stepwise cake buildup during one periodic movement of the pushing plate

A) Backstroke; B) Cake compression; C) Cake movement; D) Front dead center

a) Metal sieve; b) Cake; c) Cake built during backstroke; d) Feeding funnel; e) Pusher plate; f) Cake built during backstroke; g) Cake built at rear dead center; h) Cake built until cake movement begins; j) Cake build at front dead center

During the backstroke (Fig. 15) of the pusher plate and the change of direction at rear dead center a homogenous cake is formed. When the pusher plate advances the cake is compressed and sheared [12], but the cake still grows due to the

suspension feed at the same time. When the cake is high enough to withstand the mechanical stress the whole cake in the drum comes into movement. From that moment on the newly built cake is no longer compressed, and it moves together with the rest of the cake in the drum. The pushing force (Eq. 24) which must be imposed by the hydraulic system can be calculated from the shear stress of the cake while it slides over the sieve.

$$F_{\text{push}} = \tau_{\text{y}} \cdot 2 \cdot \pi \cdot R \cdot L \tag{24}$$

where

$$\tau_{\text{y}} = \mu \cdot \frac{\left[R^2 - (R-H)^2\right]}{2R} \cdot \rho_{\text{c}}\, gC \tag{25}$$

and for small cake heights ($R \gg H$):

$$\tau_{\text{y}} = \mu\, H \rho_{\text{c}}\, gC \tag{26}$$

The cake density (Eq. 27) is a function of the porosity and the degree of liquid saturation:

$$\rho_{\text{c}} = \rho_{\text{S}}\,(1-\varepsilon) + \rho_{\text{L}}\,\varepsilon\, S(t) \tag{27}$$

The compression of the cake during the forward movement of the pusher plate can be quantified with the conveying efficiency (Eq. 28), which is the ratio of the distance the cake is moving and the stroke of the pusher plate:

$$\xi = \frac{\Delta\, y_{\text{cake}}}{\Delta\, y_{\text{stroke}}} = \frac{\dot{V}_{\text{susp}} c_{\text{m,v}}}{\pi \left[R^2 - (R-H)^2\right] f_{\text{push}}\, \Delta\, y_{\text{stroke}}} \tag{28}$$

for $\xi < 0.85$.

Using a mass balance allows the conveying efficiency to be written as a function of the feed flow of suspension. The concentration $c_{\text{m,v}}$ (Eq. 29) can be determined from a sedimentation experiment or calculated from the concentration c_{m}:

$$c_{\text{m,v}} = \frac{1}{\left[\left(\frac{1}{c_{\text{m}}-1}\right)\frac{\rho_{\text{S}}}{\rho_{\text{L}}} + 1\right](1-\varepsilon)} \tag{29}$$

The maximum conveying efficiency is about 0.8–0.9, so a good approximation for the maximum conveying efficiency is 0.85. Therefore, at low suspension feed rate the conveying efficiency increases with increasing feed rate and the cake height remains constant (Fig. 16). The constant cake height is called the natural cake height. When the maximum conveying efficiency has been reached the cake height in the centrifuge starts to increase and the conveying efficiency remains at a constant value of about 0.85.

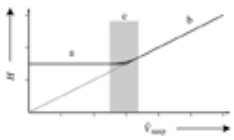

Figure 16. Cake height as a function of feed rate

a) Natural cake height; b) Increasing cake height; c) Transition zone

Simplifying the cake geometry ($H \ll R$) allows the cake height to be written as a function of the conveying efficiency:

$$H = \frac{\dot{V}_{\text{susp}} c_{\text{m,v}}}{2\,\pi\, R\, \Delta\, y_{\text{stroke}} f_{\text{push}}\, \xi} \tag{30}$$

A detailed approach to calculate the cake height from the mechanical properties of the centrifuge cake is available in the literature [1].

The solids throughput (Eq. 31) of a pusher centrifuge depends on the cake geometry, the conveying efficiency, and the pushing frequency.

$$\dot{m}_{\text{S}} = \pi \left[R^2 - (R-H)^2\right] \xi\, f_{\text{push}}\, \Delta\, y_{\text{stroke}}\, \rho_{\text{S}}\,(1-\varepsilon) \tag{31}$$

The calculation of residence time of the cake within the centrifuge is a central question concerning the dewatering and washing results (Eq. 32). Due to its conception the pusher centrifuge allows the longest residence time of all continuous filtering centrifuges, up to 15 s.

$$t_{\text{res}} = \frac{L}{\Delta\, y_{\text{stroke}}\, \xi} \cdot \frac{1}{f_{\text{push}}} \tag{32}$$

The residence time decreases with increasing conveying efficiency. At constant maximum conveying efficiency the residence time is constant (Eq. 33). This is the minimum residence time $t_{\text{res,min}}$ in the pusher centrifuge:

$$t_{\text{res,min}} = \frac{L}{\Delta\, y_{\text{stroke}} \cdot 0.85} \cdot \frac{1}{f_{\text{push}}} \tag{33}$$

The residence time in the pusher centrifuge usually is sufficient to reach approximately the Bond plateau when the centrifuge is operated with natural cake height and high residence time. When the pusher centrifuge is operated with increasing cake height the final residual moisture usually is above the equilibrium moisture indicated in the Bond diagram.

The residence time available for dewatering usually is shorter, because 30–50 % of the residence time available is needed to conduct the washing step.

4.5.2. *Washing on the Pusher Centrifuge*

Washing on continuous filtering centrifuges differs from that on discontinuous filtering centrifuges. The consumption of washing liquid is significantly lower; it amounts to 20–200 kg per tonne of dry solids. Therefore, the pore volume in the cake is not entirely filled with the washing liquid. The washing liquid drains as a film or surface flow at the particle surface.

The washing liquid is sprayed onto the cake surface through several nozzles mounted on a nozzle holder. For countercurrent washing (Fig. 17) more than one nozzle holder is required. The filtrate-collecting zone in the centrifuge body is divided into different zones. The filtrate from cake formation, the washing liquid of the first washing step, and the washing liquid of the second washing step are collected independently. The liquid of the second washing step can be recycled and sprayed onto the cake in the first washing stage.

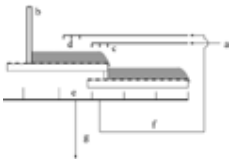

Figure 17. Scheme of countercurrent washing on a two-stage pusher centrifuge

a) Fresh washing liquid; b) Pushing plate; c) Nozzles (fresh liquid); d) Nozzles (recycled liquid); e) Partitions in the centrifuge housing to collect the filtrate and washing liquid; f) Recycle loop of washing liquid; g) Discharge of washing liquid

Most centrifuge suppliers position the nozzle holder directly after the cake-formation zone. At this position the cake has its maximum moisture; the isolated liquid bridges within the cake are not yet formed. Therefore, most of the pore liquid is accessible to the washing liquid [4], [13]. The dryer the cake, the less efficient the washing liquid.

4.5.3. *Feeding — Danger of Basket Flooding*

Feeding of the pusher centrifuge must consider that a mechanically stable cake is required in the feeding zone at the pusher plate. The necessary stability can only be given when there is no liquid pool above the newly built cake. Therefore, the operation parameters, which are mainly the solids concentration and the feed rate, must be controlled; otherwise, basket flooding occurs. At low solids concentration in the suspension, more liquid is injected to the machine than can be transported through the filter cake. In case of high feed rate the cake thickness and hence the cake resistance increase and the liquid transport of the filtrate decreases. In both cases a liquid pool at the top of the cake in the feeding zone is formed. This liquid starts to flow at the top of the cake towards the discharge side. Due to erosion effects channeling occurs during this liquid flow. The liquid flow increases the deeper the channels get, because the fed suspension also starts to run off the feeding zone. These channels quickly become deeper, causing a maldistribution of the solid in the drum. Finally the centrifuge itself comes out of balance and has to be stopped. The effect is called flooding of a pusher centrifuge.

The maximum feed rate (Eq. 34) which avoids flooding can be calculated if the specific cake resistance or the permeability of the cake is known:

$$\dot{V}_{\mathrm{susp,max}} = \frac{1}{1 - c_{\mathrm{m,v}}} \cdot \frac{\rho_{\mathrm{L}}\, Cg2\, \pi\, R\, \xi\, \Delta\, y_{\mathrm{stroke}}}{\eta_{\mathrm{L}}\, r_{\mathrm{c}}} \tag{34}$$

It can be concluded that the higher the solids concentration in the feed is, the more suspension can be fed into the pusher centrifuge. At high solids concentrations the pusher centrifuge has a high throughput. Therefore pusher centrifuges often are installed in combination with a thickening device (thickener, hydrocyclone, drum filter, DSM screen).

4.5.4. *Multistage Pusher Centrifuges*

The multistage conception has its main advantage in the reduction of the required pushing pressure. Most multistage pusher centrifuges are two-stage machines. The centrifuge drum consists of two concentric baskets. The inner basket conducts the cyclic horizontal movement. The backplane and the outer basket are fixed in the horizontal direction. The pushing force is transferred to the cake in the inner basket when the inner basket retreats. Then the wedge-wire sieve is pulled backwards under the cake, which is fixed by the contact to the stationary backplate. At the edge of the inner basket the annulus of cake falls in radial direction to the sieve of the outer basket. During the forward movement the cake in the inner basket is without motion relative to the sieve, but the front edge of the inner basket pushes the cake on the sieve of the outer basket toward the discharge side.

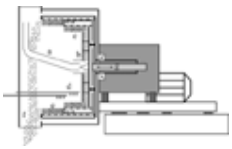

Figure 18. Concept of a two-stage pusher centrifuge

a) Feeding pipe; b) Feed accelerator; c) Cake buildup; d) Washing pipe; e) Cake at second stage; f) Cake discharge in chute; g) Hydraulic pushing system

The feeding of the multistage centrifuge occurs in the inner basket. The feeding conditions and calculations are similar to those of a single-stage centrifuge. Multistage centrifuges are less sensitive to flooding than single-stage machines, because the erosion of the cake stops at the transition to the next stage. Some three- or four-stage pusher centrifuges are

even operated in such a way that the inner basket is a thickening stage without cake formation. Nevertheless, multistage centrifuges usually also must be stopped when flooding occurs.

The pushing pressure and hence the size of the hydraulic system in the rotating hollow shaft are reduced by the two-stage conception, because of the smaller length of the cake to be transported. The pushing pressure is proportional up to the third power of the length of basket [1].

4.5.5. *Special Pusher Centrifuges*

The double-acting pusher centrifuge is a special construction with a single basket and a pushing plate in the middle position of the basket. It can be seen as a combination of two independent centrifuges in a back-to-back configuration. Both parts of the centrifuge use the same feed piping. The feed into each basket depends on the position of the pusher plate. The suspension is fed into that side of basket in which the pusher plate is in backward movement or at the backward dead center. This cyclic feed flow reduces the danger of flooding. Furthermore, the pushing plate is equipped with drainage holes through which any overlying liquid can drain to the opposite side of the basket, instead of flowing over the cake and provoking flooding. With this concept higher suspension throughput and robust operation are possible. The main disadvantages of the double-action pusher are the increased solids loss through the screen and the strong limitations in cake washing.

Pusher centrifuges, as well as conical screen centrifuges, vibrating conical screen centrifuges, and screen scroll centrifuges, use wedge-wire screens as filter media. Due to the sliding transport of the cake quite intense abrasion of the screens occurs. A second problem is sieve clogging when particles are clamped in the slot between two screen bridges. Therefore, the screen bridges are designed with a special shape. The slot widens from the cake side to the drainage basket side. This shape allows particles entering the screen slot not to be clamped, but to be transported with the filtrate. The widening at the screen slot is not constant. Near the cake side the gap has much less widening, and therefore after abrasion of the screen surface the slot width and hence the separation properties remain quite constant.

The wedge-wire screens usually consist of stainless steel and are manufactured by welding, milling, or multiple-turn coiling. The direction of the screen slots is parallel to the solids movement.

4.6. Decanting Centrifuge/Solid-Bowl Centrifuge

4.6.1. *Application for Filtration Dewatering*

Decanting centrifuges can be used for dewatering of coarse products (d_p > 50 µm) as well as for sedimentation (⟶ Centrifuges, Sedimenting). Their desaturation is a filtration process which occurs during transport of the sediment or cake in the conical part of the centrifuge. The cake is transported with the help of the screw uphill in the conical part of the solid bowl. The pore liquid drains due to the centrifugal force when the cake emerges from the liquid pool which fills the cylindrical part of the centrifuge. The residence time within the conical part of the machine is limited; it amounts to about 2 s. Because of the solid-bowl construction, the liquid flows along the bowl wall back to the pool. The limiting component of the dewatering capacity is the gap between screw and bowl.

4.6.2. *Dewatering Model*

In steady-state operation dewatering can be modeled with a simple, but efficient, approach [1]. The model is valid for coarse products, which easily reach the Bond plateau at the centrifugal values within the decanter. The model therefore does not take into account the filtration properties of the product itself. The mass balance (Eq. 35) around the control volume “conical part of the centrifuge” can be written as:

$$\dot{m}_{\mathrm{L},\uparrow} = \dot{m}_{\mathrm{L},\mathrm{R_M}} + \dot{m}_{\mathrm{L},\downarrow} = \dot{m}_{\mathrm{L},\downarrow} + \dot{m}_{\mathrm{S}} \cdot \frac{R_{\mathrm{M\,Bd}}}{1 - R_{\mathrm{M\,Bd}}} \tag{35}$$

where $\dot{m}_{\mathrm{L},\uparrow}$ is liquid transported uphill, $\dot{m}_{\mathrm{L},\downarrow}$ liquid flowing downhill in the conical part, and $\dot{m}_{\mathrm{L},\mathrm{R_M}}$ liquid adhering to solid $\dot{m}_{\mathrm{S}}$ forming the residual moisture R_{MBd}.

$$R_{\mathrm{M}} = \frac{m_{\mathrm{L}}}{m_{\mathrm{L}} + m_{\mathrm{S}}} = \frac{\dot{m}_{\mathrm{L}}}{\dot{m}_{\mathrm{L}} + \dot{m}_{\mathrm{S}}} \tag{36}$$

The liquid transported downhill can be expressed with the residual moisture (Eq. 36), where R_{Msat} is the moisture of the entirely saturated cake, when it enters the control volume (Eq. 37).

$$\dot{m}_{\mathrm{L},\downarrow} = \dot{m}_{\mathrm{S}} \cdot \left(\frac{R_{\mathrm{M_{sat}}}}{1 - R_{\mathrm{M_{sat}}}} - \frac{R_{\mathrm{M_{Bd}}}}{1 - R_{\mathrm{M_{Bd}}}} \right) \tag{37}$$

When the suspension (liquid and solid) throughput of the decanting centrifuge is increased the liquid flow downhill comes to its maximum. The maximal liquid flow (Eq. 38) can be assigned to critical solids throughput $\dot{m}_{\mathrm{S}}^{*}$. The maximal liquid flow can be written:

$$\dot{m}_{\mathrm{L},\downarrow}^{*} = \dot{m}_{\mathrm{S}}^{*} \cdot \left(\frac{R_{\mathrm{M_{sat}}}}{1 - R_{\mathrm{M_{sat}}}} - \frac{R_{\mathrm{M_{Bd}}}}{1 - R_{\mathrm{M_{Bd}}}} \right) \tag{38}$$

Any further increase of the throughput will increase the residual moisture in the solids discharge of the centrifuge. The liquid adhering to the solids in the case of $\dot{m}_{\mathrm{S}} > \dot{m}_{\mathrm{S}}^{*}$ is:

$$\dot{m}_{\mathrm{L},\mathrm{R_M}}^{*} = \dot{m}_{\mathrm{S}}^{*} \cdot \frac{R_{\mathrm{M_{Bd}}}}{1 - R_{\mathrm{M_{Bd}}}} + \left(\dot{m}_{\mathrm{S}} - \dot{m}_{\mathrm{S}}^{*} \right) \frac{R_{\mathrm{M_{sat}}}}{1 - R_{\mathrm{M_{sat}}}} \tag{39}$$

The moisture content produced by the decanting centrifuge is:

$$R_{\mathrm{M}} = R_{\mathrm{M_{Bd}}} \quad \text{when} \quad \dot{m}_{\mathrm{S}} < \dot{m}_{\mathrm{S}}^{*} \tag{40}$$

and

$$R_{\mathrm{M}} = \frac{\frac{R_{\mathrm{M_{Bd}}}}{1-R_{\mathrm{M_{Bd}}}} + \left(\frac{\dot{m}_{\mathrm{S}}}{\dot{m}_{\mathrm{S}}^{*}} - 1\right) \cdot \frac{R_{\mathrm{M_{sat}}}}{1-R_{\mathrm{M_{sat}}}}}{\frac{\dot{m}_{\mathrm{S}}}{\dot{m}_{\mathrm{S}}^{*}} + \frac{R_{\mathrm{M_{Bd}}}}{1-R_{\mathrm{M_{Bd}}}} + \left(\frac{\dot{m}_{\mathrm{S}}}{\dot{m}_{\mathrm{S}}^{*}} - 1\right) \cdot \frac{R_{\mathrm{M_{sat}}}}{1-R_{\mathrm{M_{sat}}}}} \tag{41}$$

when $\dot{m}_{\mathrm{S}} > \dot{m}_{\mathrm{S}}^{*}$

Figure 19 is a schematic plot of a typical chart and shows the constant moisture for solids throughput smaller than the critical throughput (Eq. 40). With further increasing solids throughput the moisture in the discharge steadily increases (Eq. 41). The critical solids throughput depends on the properties of the cake formed by sedimentation (porosity, equilibrium moisture at Bond plateau). This simple approach can be used for scale-up purposes [1].

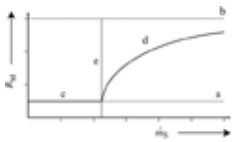

Figure 19. Critical solids throughput

a) Minimum saturation at Bond plateau; b) Maximum saturation *S* = 1; c) Constant moisture content below critical throughput; d) Increasing moisture content; e) Critical throughput

4.7. Screen-Bowl Centrifuge

The screen-bowl centrifuge can be seen as a combination of a decanting centrifuge and a screen scroll centrifuge. It has a sedimenting section, where cake formation takes place. The cake, which is transported up the conical part of the decanting centrifuge enters the second process volume, a cylindrical basket equipped with metal screens. Both drum and basket are mounted in series on the same shaft. One screw ensures solids transport in both parts of the centrifuge. Decanter drum and basket are connected to each other and rotate at the same speed. The effective diameter of the basket is larger than that of the drum; therefore, the centrifugal number in the filtering part of the centrifuge is higher as well.

In the screen-bowl centrifuge the function of thickening and the function of filtering are separated; each is carried out in an individually optimized part of the centrifuge. Due to the higher residence time in the filtering part compared to a conventional decanter or screen scroll centrifuge the screen-bowl centrifuge has improved washing capacity. Due to its conception the filtrate and the used washing liquid can be collected separately, which simplifies recycling.

[Top of Page]

5. Hybrid Processes

Hybrid processes involving filtering centrifugation are mainly a combination of mechanical dewatering and thermal drying and aim at a dryer product with better handling properties. Some strategies line up several process steps, like the centrifugal dryer, which are conducted within the centrifuge basket. Real hybrid processes try to use the synergy effects which derive from the combination of mechanical and thermal effects.

5.1. Hyperbaric Centrifugation/Jet-Stream Centrifugation

The combination of gas differential pressure filtration and centrifugal filtration is called hyperbaric [14], [15] or jet-stream centrifugation. It can be operated in specially designed filtering centrifuges such as the siphon peeler centrifuge $(\Delta p_{\max} = 1\ \mathrm{bar})$ or the inverting filter centrifuge $(\Delta p_{\max} = 6\ \mathrm{bar})$. The additional driving force of the gas pressure or the vacuum accelerates liquid drainage during the plug-flow phase (Eq. 43). The gas pressure only slightly influences the film flow.

$$\dot{V} = \frac{\pi\, L_{\mathrm{dr}} [\, \rho_{\mathrm{L}}\ \omega^2\, (R^2 - r_{\mathrm{L}}{}^2) + 2\, \Delta\, p]}{\eta_{\mathrm{L}}\ r_{\mathrm{C}} \ln \frac{R}{r_{\mathrm{S}}}} \tag{42}$$

$$\frac{1}{8\,\lambda} = \frac{\Delta\, t\, \rho_{\mathrm{L}}\, g C d_{\mathrm{h}}{}^2}{32\, \eta\, H} = \left(1 - \frac{x}{H}\right) + \frac{\Delta\, p - p_{\mathrm{ke}}}{\rho_{\mathrm{L}}\, g C H} \cdot \ln \left[\frac{\frac{x}{H} + \frac{\Delta p - p_{\mathrm{ke}}}{\rho_{\mathrm{L}} g C H}}{1 + \frac{\Delta p - p_{\mathrm{ke}}}{\rho_{\mathrm{L}} g C H}}\right] \tag{43}$$

The superimposed gas pressure is an additive term in the known equations for dewatering kinetics using the kinetic parameter λ and the filtrate flux (Eq. 42).

The field of application of hyperbaric centrifugation primarily is in fine products, which show desaturation behavior in the Bond II regime. The gas pressure is applied to remove the remaining liquid in the capillaries near the filter medium. In this hybrid process the advantages of both processes are combined: solids handling and washing in the centrifugal field and the higher driving potential of the gas pressure.

5.2. Steam-Enhanced Centrifugation/Steam-Pressure Centrifugation

Instead of pressurized air, pressurized steam is used to build up the pressure difference. The process can be seen as a combination of steam-pressure filtration [16], [17] and centrifugal filtration. Steam filtration uses the positive effect of the self-stabilizing displacement front, which strongly reduces early gas breakthrough. The use of steam in steam-pressure filtration improves the dewatering result by about 20 % of saturation compared to conventional filtration with identical process parameters. Steam-enhanced centrifugation can be applied with two different strategies, either in the Bond II or in the Bond III region. In the Bond II region the pressurized steam is used for improved filtration of the remaining liquid in the capillaries. The moisture content is reduced and handling properties of the solids improved. The application of steam-enhanced centrifugation in Bond III under the condition that the saturation is less than 0.15 leads to an entirely dry product [18]. Mechanical separation and drying are performed simultaneously in same process volume. Steam-pressure centrifugation belongs to the mechanical-thermal dewatering processes.

The dewatering theory of conventional and hyperbaric centrifugation can be further extended to describe displacement with a condensable gas [19]. The plug flow is written with an energy parameter Ψ which contains the ratio of sensible to latent heat.

$$\frac{1}{8\,\lambda} = \frac{\Delta\, t\, \rho_L\, g C d_h^{\,2}}{32\, \eta\, H} = (1-\psi)\cdot\left(1-\frac{x}{H}\right) + \left[\psi - (1-\psi)\cdot\frac{\Delta\, p - p_{ke}}{\rho_L\, g C H}\right]\cdot \ln\left[\frac{\frac{x}{H}+\frac{\Delta p - p_{ke}}{\rho_L g C H}}{1+\frac{\Delta p - p_{ke}}{\rho_L g C H}}\right] \tag{44}$$

where

$$\psi = \frac{\eta_{st}}{\eta_L}\cdot\frac{c_{p,S}\,\rho_S\,\Delta\, T}{\rho_{st}\,\Delta\, h_v}\cdot\frac{1-\varepsilon}{\varepsilon} \tag{45}$$

The calculated breakthrough time (Eq. 44) is essential for economic operation of the process. Steam injection is efficient during plug flow until steam breakthrough. Any further steaming creates costs without any reduction in moisture.

The steam, which partly condenses within the pores of the cake, has a certain washing efficiency. Steaming can approximately replace an amount of washing liquid equal to the amount of condensate produced [20].

[Top of Page]

6. Laboratory-Scale Testing

The dimensioning of technical centrifugation processes requires laboratory- and pilot-scale testing. Since ca. 1995 the quality of the laboratory scale experiments and equipment has been improved. The main focus of laboratory-scale experiments is on desaturation kinetics and the equilibrium of dewatering (Bond diagram).

Effects of cake formation and cake structure require larger experimental setups because the movement of the liquid pool relative to the cake surface has a governing influence.

6.1. Laboratory Centrifuges

The standard laboratory-scale centrifuge is a bucket centrifuge with either vertical or horizontal shaft. The test bucket is filled outside of the centrifuge with a previously built cake or with suspension. The direct feeding of bucket centrifuge with suspension has the disadvantage that, due to the Coriolis force, an inclined cake surface is formed.

The acceleration time until maximum rotary speed is a criterion for division into high- and low-end laboratory-scale centrifuges. High-end laboratory centrifuges are necessary to follow the time-dependent processes within continuous centrifuges, which have a time frame of several seconds. During this interval the cake in the laboratory-scale centrifuge must be accelerated, processed, and decelerated. If an on-line sensor system is used, the deceleration is no longer relevant to the measurement.

The typical bucket concept applying an on-line sensor consists of a bucket with a filtering bottom (Fig. 20). The filtrate is collected below in a reservoir, which is equipped with a pressure sensor. This sensor has the function of a balance; it is able to quantify the mass of filtrate. The sensor signal is recorded either by telemetry, via a collector ring, or with help of a data logger. The typical acceleration time of high-end laboratory centrifuges is <thinsp;2 s. The typical centrifugal number *C* of a high-end laboratory centrifuge is about 4000 (up to 6000). With the help of structural information on the cake the time-dependent saturation can be calculated from the recorded function “filtrate mass over process time”.

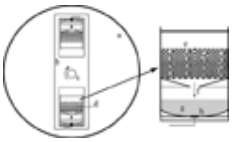

Figure 20. Conception of a high-end lab scale centrifuge and bucket

a) Housing; b) Rotor; c) Shaft; d) Bucket inserted into the rotor; e) Cake built up in the bucket; f) Filtrate funnel; g) Reservoir with collected filtrate; h) Pressure sensor for on-line measurements

The low-end laboratory centrifuges are used to estimate the equilibrium of dewatering or Bond curve. The cake is processed for several minutes and the moisture content is measured off-line. This set of information can be sufficient for the design of discontinuous processes.

An alternative to the bucket centrifuge is the so-called ring centrifuge, which is a smaller basket centrifuge. The basket length is reduced to an annulus with an opening of several centimeters. The bottom of the ring can be equipped with either a metal sieve or a filter medium. This type of centrifuge requires a larger quantity of suspension. The main focus of the investigations is on cake formation and cake washing. On-line sampling of the filtrate is possible, but it shows a quite high dead time until the filtrate creates a signal on the balance.

[Top of Page]

7. Applications

The field of application is too wide to be discussed here. Application and process integration of filtering centrifuges are described under other keywords

pharmaceutical chemicals, bulk polymers (⟶ Polyolefins – Polypropylene, ⟶ Polymerization Processes – Polymerization Processes and Reactor Modeling), minerals (⟶ Potassium Compounds – Potassium Chloride, ⟶ Coal – Coal Preparation; ⟶ Sodium Sulfates), food (⟶ Sugar – Separation of Sugar from Massecuite), fibers (⟶ Paper and Pulp – Pulp), proteins (⟶ Starch – Starch), ores, bulk chemicals, fine chemicals and intermediate chemicals (⟶ Terephthalic Acid, Dimethyl Terephthalate, and Isophthalic Acid – Production; ⟶ Adipic Acid – Production), commodity chemicals (⟶ Laundry Detergents – Production of Powder Detergents), catalysts (⟶ Heterogeneous Catalysis and Solid Catalysts – Production of Heterogeneous Catalysts).

[Top of Page]

References

Cited References

1. W. Stahl: *Industrie-Zentrifugen*, 1st ed., DrM Press, Männedorf 2004.
2. K. Luckert: *Handbuch der mechanischen Fest-Flüssig-Trennung*, 1st ed., Vulkan Verlag, Essen 2004.
3. W. Leung: *Industrial Centrifugation Technology*, McGraw-Hill, New York 1998.
4. B. Fuchs, U. A. Peuker, W. Stahl: “The purification of a porous system in the centrifugal field”, 9. WFC, New Orleans 2004.
5. U. A. Peuker: *Über die kombinierte Dampfdruck- und Zentrifugalentfeuchtung von Filterkuchen*, PhD-Thesis, Universität Karlsruhe, 2002.
6. L. Svarovsky: *Solid-Liquid Separation*, 3rd ed., Butterworths, Sevenoaks 1990.
7. R. Kessler, U. Esser, M. vom Felde, T. Schäfer: “Filter centrifuges, development and Implementation of a Model Based Control-System”, *Proc. Filtech Europe*, Düsseldorf 2003.
8. R. J. Wakeman, F. Deshun: “The Control Ring Centrifuge - a New Type of Conical Basket Centrifuge”, *Chem. Eng. Res. Des.* **69** (1991) 403–408. Links
9. F. Deshun, R. J. Wakeman, H. Zhong: “An Analysis of Gyratory Forces and Wobble Angles in Tumbler Centrifuges”, *Chem. Eng. Res. Des.* **69** (1991) 409–416. Links
10. M. Thomas, C. Veal, J. Smitham: “Novel Basket Design for Fine Coal Centrifuges”, *J. S. African Inst. Min. Metall.* **101** (2001) 47–52.
11. M. Bentz, O. Hartmetz, W. Stahl: “Particle Destruction Effects in Pusher Centrifuges during Solid-Liquid Separation”, *Chem. Eng. Technol.* **27** (2004) 1075–1080.
12. O. Imhof, Rechnerische und experimentelle Untersuchungen über die Kuchenbildung und die Filtrationsabläufe in einer einstufigen Schubzentrifuge, PhD-Thesis, Universität Stuttgart, 1979.
13. B. Hoffner, B. Fuchs, A. Heuser: “Washing Processes for Disperse Particulate Systems — Process Spectrum and Aspects for the Process Choice”, *Chem. Eng. Technol.* **27** (2004) 1065–1071.
14. C. Stadager, Die Entfeuchtung von Filterkuchen durch die Kombination von Zentrifugal- und Gasdruckkraft, PhD-Thesis, Universität Karlsruhe, 1995.
15. C. Stadager: “Gas Throughput and Thermal Drying in Hyperbaric Centrifugation”, *Chem. Ing. Tech.* **68** (1996) 1469–1473.
16. U. Peuker, W. Stahl: “Scale-up and Operation of a Steam Pressure Filter in Pilot Scale”, *Chem. Eng. Technol.* **24** (2001) 612–616. Links
17. U. A. Peuker, W. Stahl: “Steam Pressure Filtration: Mechanical-Thermal Dewatering Process”, *Drying Technol.* **19** (2001) 807–848.
18. U. A. Peuker, W. Stahl: “Flow-Through Drying of Preheated Beds of Bulk Solids Drying Phase during Mechanical-Thermal Demoisturing Processes”, *Chem. Ing. Tech.* **74** (2002) 934–950.
19. U. A. Peuker: “Extension of the model concept of centrifugal drying - Consideration of mechanical/thermal drying”, *Chem. Ing. Tech.* **75** (2003) 1237–1248.
20. B. Fuchs, U. A. Peuker, W. Stahl: “Saturated Steam — A Good Washing Medium?” Filtech Europe, Düsseldorf 2003.

Specific References

21. G. Hultsch, H. Wilkesmann: “Filtering Centrifuges,” in D. B. Purchas (ed.): *Solid–Liquid Separation*, Upland Press, Croydon 1977, pp. 493–559.
22. H. Trawinski: “Die äquivalente Klärfläche von Zentrifugen”, *Chem. Ztg.* **83** (1959) 606–612.
23. W. Bender, Zur Berechnung des Durchsatzes von Schälschleudern, Dissertation, Technische Universität Karlsruhe 1971.
24. H. Trawinski: “Zentrifugen, Trenngeräte mit höchster Abscheidungswirkung”, *Chem. Ing. Tech.* **26** (1954) 189–201.

25. F. Damrat: “Trennung heterogener Stoffe durch Zentrifugieren”, *Konstruktion* **12** (1960) 447–456.
26. W. J. Sokolov: *Moderne Industriezentrifugen*, 2nd ed., Maschinostrojenije, Moskwa 1967, and VEB Verlag Technik, Berlin 1971.
27. W. Batel: “Menge und Verhalten der Zwichenraumflüssigkeit in körnigen Stoffen”, *Chem. Ing. Tech.* **33** (1961) 541–547.
28. W. Heckmann: “Die physikalischen Grundlagen der mechanischen Feststoff-Flüssigkeits-Trennung und ihr Einfluß auf den Bau von Entwässerungsvorrichtungen”, *Glueckauf* **96** (1960) 1281–1287.
29. H. H. Hülsen: “Kontinuierliche Zentrifugen zur Trennung fest-flüssig in der chemischen Industrie”, *Verfahrenstechnik (Mainz)* **6** (1972) no. 1, 1.
30. H. Quetsch: “Neue Siebzentrifuge für Fest-Flüssig-Trennung”, *Chem. Ing. Tech.* **42** (1970) no. 4, 183–184.
31. E. V. Szantho, H. Mathiak: “Entwicklung der Schwingsiebschleuder und Einsatzmöglichkeiten zum Entwässern von aufbereitenden Produkten”, *VDI Z.* **108** (1966) 493–499.
32. O. Molerus, K. Brunner: “Verfahrenstechnische Auslegung von Schwingsiebschleudern”, *Chem. Ing. Tech.* **48** (1976) no. 2.
33. G. Fengler: “Siebschnecken-Zentrifugen”, *Chem.-Anlagen + Verfahren* (1984) nos. 10 and 11.
34. H. Trawinski: “Der neueste Entwicklungsstand auf dem Gebiet der Zentrifugen- und Hydrozyklontechnik”, *Aufbereit. Tech.* **8** (1967) 470–483.
35. G. Hultsch: “Einige Kriterien von Massengutschleudern und Beschreibung einer neuen Siebzentrifuge”, *Aufbereit. Tech.* **7** (1966) 720–723.
36. H. Wilkesmann: “Kinematik und verfahrenstechnische Optimierung einer Massengut-Zentrifuge mit Präzessionsaustrag (Taumelzentrifuge)”, *Aufbereit. Tech.* **19** (1978) 587–592.
37. E. Rüegg: “Kontinuierliche Zentrifugen für die Zuckerindustrie”, *Zucker* **5** (1952) 548–551.
38. W. Schnurrenberger: “50 Jahre Schubzentrifugen, System Escher Wyss”, *Techn. Rundsch. Sulzer* **67** (1985) no. 2.
39. H. Zürrer: “Waschprozeß in mehrstufigen Schubzentrifugen”, *Escher Wyss Mitt.* (1973) no. 2.
40. G. Hultsch, J. Wangermann: “Eine neuartige Zentrifuge zur Filtration und gleichzeitigen Abtrennung von Zwickelkapillarflüssigkeit”, *Aufbereit. Tech.* **9** (1968) 487–490.

[Top of Page]

Symbols

***A*:** area

C: centrifugal number

***c*:** concentration

$c_{m,v}$**:** sediment concentration

c_p**:** heat capacity

***d*:** diameter

***F*:** force

***f*:** frequency

f_r**:** filtration ratio

***g*:** gravity

H: cake height

H_{fl}**:** flank lead of transporting screw

Δh_v**:** latent heat

***L*:** length

***m*:** mass

$\dot{m}$**:** mass flow

***n*:** rotary speed

***p*:** pressure

p_{ke}**:** capillary pressure

***R*:** radius of drum or basket

***r*:** radius

r_c**:** specific cake resistance

***S*:** saturation

***T*:** temperature

***t*:** time

***V*:** volume

$\dot{V}$**:** volume flux

R_M**:** residual moisture

***x*:** liquid height in the cake

***y*:** distance

Greek symbols
α**:** angle (inclination of conical part of centrifuge)

δ**:** wetting angle

ε**:** porosity

η**:** dynamic viscosity

ρ**:** density

σ**:** surface tension

ϕ**:** energy parameter

τ_y**:** yield stress in direction of y

ω**:** angular speed

ζ**:** conveying efficiency

Subscripts
Bd: at Bond plateau (equilibrium)

c: cake

dr: drum

F: film

h: hydraulic

L: liquid

m: per unit mass

max: maximum

min: minimum

non: nonremovable

push: pushing

res: residence

S: solid

sat: saturated state

scr: screw

st: steam

susp: suspension

v: void

v: per unit volume

[Top of Page]

Related Articles

See also ⟶ Solid–Liquid Separation, Introduction; Centrifuges, Sedimenting; Filtration

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Centrifuges, Sedimenting

Hans Axelsson, Alfa Laval Separation AB, Tumba, Sweden
Bent Madsen, Alfa Laval Separation A/S, Søborg, Denmark

1. Introduction 2
2. Fundamentals 2
2.1. Centrifugal Sedimentation 2
2.2. Centrifugal Sedimentation with Continuous Flow 3
3. Centrifugal Separators with Continuous Feed 5
3.1. General 5
3.2. Disk-Bowl Machines 5
3.3. Chamber and Multichamber Bowls . 11
3.4. Tubular Bowl Centrifuges 12
3.5. Decanter Centrifuges 13
3.6. Other Bowl Geometries 16
4. Batch Centrifuges 16
4.1. General 16
4.2. Bottle Centrifuges 16
4.3. Batch Zonal Centrifuges 17
5. Fluid Dynamics in Centrifuges 17
5.1. Chamber and Tubular Bowl Centrifuges 18
5.2. Disk-Bowl Centrifuges 18
5.3. Decanter Centrifuges 19
6. Centrifuge Technology 21
6.1. Rotor Dynamics 21
6.2. Materials Selection 21
7. Centrifuge Selection 22
7.1. Introduction 22
7.2. Main Process Parameters 22
7.3. Test Tube Centrifugation (Spin Test) . 23
7.4. Important Properties of the Process Streams 23
7.5. Improvement of Separability 25
7.6. Pilot Machine Testing 25
7.7. Scale-Up 25
8. Applications 26
9. References 27

Symbols

a	area, m^2
d	diameter, m
D	inner diameter of test tube, m
D	inner diameter of decanter bowl, m
g	acceleration due to gravity (9.81 m/s^2)
k	constant
KQ	semi-empirical measure of centrifuge size (Eq. 15)
L	length of decanter bowl, m
n	bowl speed, min^{-1}
N	number of disks
p	pressure, Pa
Q	flow rate, m^3/s
r	radius, m
R	outer radius of rotor, m
s	settling distance, m
t	distance between disks, m
T	spin time, s
u	velocity, m/s
V	bowl volume, m^3
z	number of nozzles
Z	acceleration factor or g number
α	half-cone angle of disk
β	angle of tube in test-tube centrifuge
χ	angle of nozzle to the tangent of bowl
η	dynamic viscosity of fluid, mPa · s
λ	dimensionless number (Eq. 18)
μ	separation efficiency factor (Eq. 13)
ν	kinematic viscosity of fluid, m^2/s
ϱ	density, kg/m^3
ω	angular velocity, rad/s
Σ	area equivalent (Eq. 10), m^2

Subscripts

act	actual
c	centrifugal
d	drag
e	equivalent
f	fluid
g	gravity
i	interface
in	inner
h	heavy liquid
l	light liquid
N	nozzle

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.c05_c02.pub2

out	outer, outlet
p	particle
theor	theoretical
1	inner
2	outer

1. Introduction

Sedimenting centrifuges are used for the separation of two or three phases: liquid – liquid, liquid – solid and liquid – liquid – solid. Sedimenting centrifuges use centrifugal force to accelerate the sedimentation process. By rotating the process fluid, the sedimentation rate can be increased by a factor of several thousands compared to static sedimentation or settling. Other types of equipment also use centrifugal force for phase separation. This article deals only with sedimenting centrifuges. For filtration centrifuges and hydrocyclones, see → Hydrocyclones and → Centrifuges, Filtering.

In *liquid – liquid separation* the most common alternative to sedimenting centrifuges are static settlers. The advantage of the centrifugal separator compared to the settler is the small equipment volume and/or short residence time.

The choice between a centrifugal filter and a sedimenting centrifuge in *solid – liquid separation* depends on the particle size and the permeability of the cake of solids in the centrifugal field [17].

Cross-flow filtration may be an alternative to centrifugation for small particle sizes and mostly for small flow volumes.

2. Fundamentals

2.1. Centrifugal Sedimentation

General. The sedimentation rate of solid particles or droplets in the gravity field is a function of the particle (or droplet) size, the density difference, and the viscosity of the suspension (or emulsion). In applications in which centrifuges are used, the static sedimentation velocity is between 10^{-9} and 10^{-4} m/s. To accelerate the process, centrifuges apply centrifugal forces of 500 – 30 000 g.

At lower sedimentation rates, special ultracentrifuges with up to 900 000 g are used to separate macromolecules, virus particles, and microbial cell constituents. Sedimentation velocities are expressed as sedimentation coefficients in Svedberg units (S), have the dimension seconds, and are defined as 10^{13} times the settling velocity per unit acceleration (in m/s^2). They range from 4 to 10^6 S [5].

A third class of centrifugal separators are gas centrifuges, used primarily to separate isotopes (→ Isotopes, Natural, Chap. 4.2.3). Uranium enrichment is the best known application. The centrifugal force is greater than 150 000 g. The enrichment in a single centrifuge is very small, so a large number of machines have to be arranged in series to obtain a sufficient degree of enrichment.

Settling Velocity. The settling velocity of the particles or droplets is a fundamental property of a suspension or emulsion and determines the flow rate through or size of the centrifuge together with the residence time and settling length. In most centrifugations the settling velocity u_c is so small that Stokes law applies:

$$u_c = \frac{(\varrho_p - \varrho_l)\, d_p^2}{18\,\eta} \cdot r\,\omega^2 \ [\mathrm{m/s}] \tag{1}$$

where ϱ_p and ϱ_l are the densities of particle and surrounding medium, d_p is the particle diameter, η is the viscosity of the suspension, r is the distance from the center of rotation, and ω is the angular velocity, which is related to the bowl speed n in rpm according to the relation

$$\omega = \frac{2\,\pi\,n}{60} \ [\mathrm{s}^{-1}] \tag{2}$$

The acceleration factor Z (the g number) is

$$Z = \frac{r\,\omega^2}{g} \tag{3}$$

where g is the acceleration due to gravity. The settling velocity u_g in the gravity field is

$$u_g = u_c \cdot \frac{g}{r\,\omega^2} \ [\mathrm{s/m}] \tag{4}$$

The Equations (1) and (4) are valid under the conditions that

1) The flow round the particle is laminar (Reynolds number $Re < 0.4$)
2) The settling of a particle is not influenced by other particles, i.e., the concentration is very low

For particle $Re > 0.4$ alternative equations have been derived [2, 5]. For high particle concentrations, a phenomenon called hindered settling sets in, decreasing the settling rate. Equations for correction factors < 1 for the settling rates u_c and u_g in Equations (1) and (4) are summarized in [9]. At 20 vol % of particles, the settling rates decrease by a factor > 2.

Residence Time and Settling Length. A high g number (Z) is not the only factor that determines the flow rate at acceptable separation efficiency through a centrifuge. The residence time in the centrifugal force field, i.e., the rotor volume, can compensate for a small Z. A small settling length, i.e., the distance a particle must travel to sbe captured, will also improve the flow rate at acceptable separation efficiency.

The Bottle Centrifuge. The settling velocity of particles that are separated in a centrifuge is often so small that its measurement is not practical without application of centrifugal force. The bottle (swing-out) centrifuge, shown schematically in Figure 1, is often a suitable tool for measurement of the settling velocity.

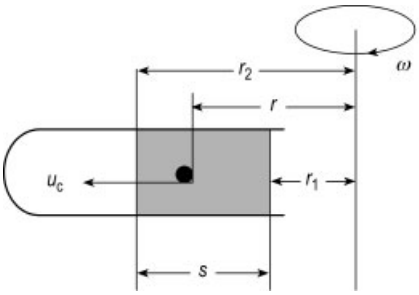

Figure 1. Bottle swing-out centrifuge

The following equations describe the particle motion:

$$u_c = \frac{dr}{dt} \tag{5}$$

$$u_c = u_g \cdot \frac{r\,\omega^2}{g} \tag{6}$$

The aim is for all particles to move to a radius $> r_2$. In the worst case, the particle starts at the liquid surface at r_1. Insertion of Equation (6) into Equation (5), rearranging, and integrating gives

$$\int_0^T dt = \frac{g}{u_g\,\omega^2} \int_{r_2}^{r_1} \frac{dr}{r} \tag{7}$$

$$T = \frac{g}{u_g\,\omega^2} \cdot \ln \frac{r_2}{r_1} \tag{8}$$

where T is the spinning time necessary for all particles to leave the volume between levels r_1 and r_2. By rearranging Equation (8) one obtains

$$u_g = \frac{g}{T\,\omega^2} \cdot \ln \frac{r_2}{r_1} \tag{9}$$

It is clear that a small settling velocity necessitates a long spinning time, i.e., residence time, and/or a short sedimentation distance at a high angular velocity. This is also valid for centrifuges with continuous feed and effluent flow. Because of mechanical and hydraulic limitations, these requirements are in conflict with each other and can not be met in industrial centrifuges without compromises. This has led to the development of a broad range of machines in which different requirements have been given priority.

2.2. Centrifugal Sedimentation with Continuous Flow

The Σ Theory.

The Generalized Σ Formula. The most used quantity to characterize centrifuges, the Σ concept, was presented by Ambler [36, 37]. It is the calculated equivalent surface area of a static settling tank with the same theoretical performance. In its derivation he considered particles with a critical diameter d_c that were separated to 50 %. Today, however, the most widely used definition of the critical particle is that which is separated to 100 %. This does not influence the formulas for the Σ value, but the value for the feed flow rate Q is halved.

Σ is defined as

$$Q_{theor} = u_g\,\Sigma \tag{10}$$

where u_g is the Stokes settling velocity (Eq. 1). Σ is given by the general expression

$$\Sigma = \frac{V}{s_e} \cdot \frac{\omega^2 \cdot r_e}{g} \tag{11}$$

where V is the liquid volume in the bowl, s_e is the effective settling distance, and r_e is the effective distance from the center of rotation.

With Equations (1) and (4) the equation for the critical diameter becomes

$$d_c = \left(\frac{18\,\eta\,Q_{\text{theor}}}{(\varrho_p - \varrho_f)} \cdot \frac{s_e}{V \cdot \omega^2 \cdot r_e}\right)^{\frac{1}{2}} = \left(\frac{18\,\eta\,Q_{\text{theor}}}{\Sigma\,(\varrho_p - \varrho_f)\,g}\right)^{\frac{1}{2}} \tag{12}$$

The derivation above is based on the assumption that

- Viscous drag determines the particle movement
- The flow in disk bowls between the disks is laminar and symmetrical
- The liquid rotates at the same speed as the bowl
- The particle concentration is low (no hindered settling)
- The particle at all times moves with its final settling velocity u_c
- The settling velocity u_c is proportional to the g force.

Ambler uses the critical particle in his formula; therefore, his analysis is a special case of Svarovsky's grade efficiency function [7], in which for each particle size the ratio of sedimented mass to the mass in the feed material is determined and integrated over all particle sizes to give a more realistic measure of the separation efficiency. The terminology, definitions and theory of particle classification is dealt with comprehensively in [9].

Σ for Various Centrifuges. The flow system in between two conical disks in a disk-bowl centrifuge is shown in Figure 24. The particle has a radial velocity component u_c because of the centrifugal force, which increases with increasing r. It has also a drag dependent velocity component u_d which can have any size or direction depending on the flow situation round the particle. In the Σ theory it is assumed that it depends on uniform radial plug flow parallel to the disks. The derivation of the Σ for various types of centrifuge is found in [37], and the results are summarized in the following:

Disk-bowl [36, 37]

$$\frac{\pi \cdot \omega^2}{g} \cdot \frac{2}{3} \cdot N \cdot \left(r_2^3 - r_1^3\right) \cdot \cot\alpha$$

r_2 = max. radius of disk,
r_1 = min. radius of disk,
N = number of disk,
α = half-cone angle of disk

Decanter [7, 37]

$$\frac{\pi \cdot \omega^2}{g}\left[L_1\left(\frac{3}{2} \cdot r_2^2 + \frac{1}{2} \cdot r_1^2\right) + L_2\left(\frac{r_2^2 + 3\,r_2\,r_1 + 4\,r_1^2}{4}\right)\right]$$

L_1 = length of cylindrical part,
L_2 = length of conical part,
r_1 = inner radius of liquid,
r_2 = inner radius of bowl,

Tubular-bowl, chamber-bowl [37]

$$\frac{\pi \cdot \omega^2}{g}\,L\,\frac{r_2^2 - r_1^2}{\ln\left(\frac{2\,r_2^2}{r_2^2 + r_1^2}\right)}$$

r_1 = inner radius of liquid,
r_2 = inner radius of bowl,
L = inner length of bowl,

Multichamber-bowl [2]

$$\frac{\pi \cdot \omega^2}{g}\,\frac{L}{3}\sum_{i=0}^{i=n}\frac{r_{2\,i+1}^3 - r_{2\,i+2}^3}{r_{2\,i+1} - r_{2\,i+2}}$$

indices with even numbers = inner radius of chamber,
indices with odd numbers = outer radius of chamber,
$n + 1$ = number of chambers,
L = height of chambers

Separation Efficiency. The assumptions in the Σ theory are not fulfilled in reality. Therefore, Equation (10) should include an efficiency factor μ, so that in practice the flow rate is lowered for the required separation performance. Estimated efficiency factors are listed in the following [6]:

Disk-bowl machines	45 – 73 %
Decanter centrifuges	54 – 67 %
Tubular bowls	90 – 98 %

By careful measurement of the density difference and particle dimensions in one case an efficiency factor of max. 16 % was found to fit the experimental data in a pilot-scale disk bowl machine [38]. However, this probably also included effects of hindered settling.

In the absence of hindered settling, the equation describing the relationship between the feed flow rate, the Σ value, and the settling rate, is thus

$$Q_{\mathrm{act}} = \mu\, u_{\mathrm{g}}\, \Sigma \tag{13}$$

The KQ Formula for Disk-Bowl Centrifuges. A semi-empirical equation for the "separation area" [39] is based on the finding that separation data fit the relation

$$\frac{u_{\mathrm{c}}}{u_{\mathrm{g}}} = \left(\frac{\omega^2\, r}{g}\right)^k \tag{14}$$

with $k = 0.75$ better than the classical analysis (Eq. 4) with $k = 1$. This is a consequence of the increasing shear forces in the Ekman layers (see Section 5.2). For practical purposes the following relation is used:

$$\mathrm{KQ} = 280 \cdot \left(\frac{n}{1000}\right)^{1.5} \cdot N \cot\alpha\, \left(r_{\mathrm{out}}^{2.75} - r_{\mathrm{in}}^{2.75}\right) \tag{15}$$

where n is the bowl speed in rpm and r_{out} and r_{in} are the radii of the disk in centimeters. This form of the KQ formula is dimensionally incorrect, but it is mathematically possible to make it correct and to express the KQ value as an entity with the dimension of area. In practice the difference between KQ and Σ for scale-up purposes is negligible, compared to other inaccuracies.

Hindered Settling. One of the few experimental investigations into hindered settling of cells [38] shows that complex phenomena occur at high concentrations. A model with several experimental constants was developed. It showed that settling rates decrease rapidly with increasing solids contents, by a factor of about 2 at 20 vol % and by 15 – 20 at 50 vol %. The classical model used in [39] gave similar results. The decreased settling velocity can be attributed to increased viscosity of the suspension at high solids concentrations [9]. In the tests in [38] it was also found that cells may settle as flocs at higher concentration. Hindered settling is one further example of the difficulty to calculate theoretically the performance of centrifuges.

Gas Centrifuges. The theory for gas centrifugation is too complicated to be included here. The performance of a gas centrifuge is, besides peripheral velocity and rotor length, very much depending on the flow pattern in the rotor and the location of entry and exit points of the streams. The reader is referred to [5] and [13].

3. Centrifugal Separators with Continuous Feed

3.1. General

The different types of continuous centrifuge were developed from different principles of solids handling. The first centrifuges were developed for separating a process fluid (e.g., milk, fermented yeast broth) into two fluid streams. Heavy solids that could not flow out were collected at the periphery and removed by periodic manual cleaning. These were followed by centrifuges with peripheral nozzles that discharge solids as a slurry. In many processes these replaced machines needing expensive manual cleaning. For processes in which the solids must be collected in a more concentrated, nonfluid, form, machines with internal conveyors (decanter centrifuges) were developed. Intermittently discharging machines were developed to fill the gap between the three above-mentioned types and are now the most common continuous centrifuges. A number of discharging mechanisms have been developed over the years. A bottle centrifuge with continuous feed and separated effluent (centrate) is available.

The most comprehensive description of centrifuges can be found in [2]. Brochures and technical publications from centrifuge manufacturers provide the most modern information about machine types, sizes, and capacities.

3.2. Disk-Bowl Machines

General. In disk-bowl machines the sedimentation path is made short by introduction of conical discs, separated by spacers. The distance is between 0.4 and 3 mm, depending on the viscosity and solids content. The half-cone angle is in the range 35 – 45° and is smaller for solids with a small angle of repose.

The spacers between the disks have different designs, depending on the separation duty. The most common spacers are long ribs placed radially along the cone; these help the liquid retain its rotation velocity. The majority of disk-bowl machines are used for separating two liquid phases. In many cases, droplets of the dispersed phase in the feed liquid coalesce atan interface inside the bowl, and the newly formed liquid leavesthe

bowl continuously through a second outlet, separated from the original continuous phase. The position of the interface in the bowl is important and can be controlled by, for example, adjusting the radius of one of the weirs, usually that for the heavy phase. Its radius r_h (see Fig. 2) is given by

$$r_h = \sqrt{r_i^2 \left(1 - \frac{\varrho_l}{\varrho_h}\right) + r_l^2 \cdot \frac{\varrho_l}{\varrho_h}} \tag{16}$$

where r_i is the radius of the interface, r_l the outlet radius of the light liquid phase, and ϱ_l and ϱ_h are the densities of the light and heavy liquid phases. The equation is derived from a pressure balance but is a simplification because it neglects pressure drops and weir heights. It is adequate for a first approximation. It is also applicable to chamber bowls, tube centrifuges and decanters (Sections 3.3, 3.4 and 3.5).

If the objective of the separation is to obtain a very clean light phase, the interface should be placed at a large radius to avoid heavy droplets' contaminating the light phase. For obtaining a clean heavy phase, the interface should have a small radius.

Several important applications of liquid – liquid separations in disk bowls do not involve coalescence at an interface. The most common examples are separation of cream from milk and alkali refining of vegetable oil. The split between the outlet streams is then controlled by counterpressures that can be applied in several ways.

Solid-Bowl Centrifuge. Figure 2 shows a solid-bowl centrifuge requiring manual cleaning. Heavy solids are collected at the periphery; the machine must be dismantled and the bowl opened to remove the solids, often by lifting up a collecting basket, fitted inside the bowl. The feed enters the bowl in the center through a still-standing pipe. The version of bowl shown is suitable for liquid – liquid – solid separation and therefore has two liquid outlets. The light liquid flows through the disk stack and out of the bowl through a still-standing centripetal pump known as a paring disk. The heavy liquid passes over a disk near the bowl hood and is removed through another paring disk. The paring disk converts the kinetic energy of the rotating liquid, and discharges it under pressure.

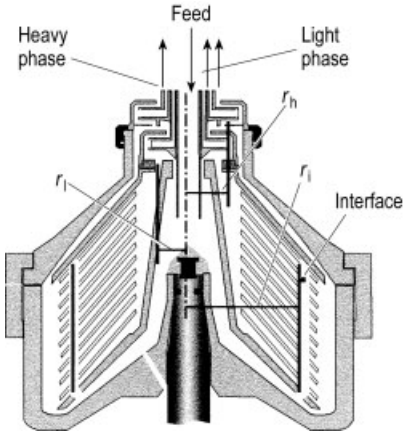

Figure 2. Solid bowl with disks, requiring manual removal of solids (vertical section)

This type of bowl has diameters between 140 and 750 mm. The solids space has a volume of up to 34 L. Further data are listed in Table 1. The principles illustrated in Figure 2 are also applicable to other bowl types. Instead of paring disk(s), open outlets over weirs can be used. In many cases only one of the two outlets has a paring disk.

A small solid-bowl machine is shown in Figure 3.

Figure 3. A small solid-bowl centrifuge (type Emmie; courtesy Alfa Laval Separation AB)
The bowl diameter is 140 mm.

Table 1. Characteristics of continuous-feed centrifuges [a]

Type	Mode of solids discharge	Feed flow rate, L/h	Feed solids content, vol %	Solids flow rate, L/h	Σ value, m^2	Max. Z developed	Consistency of solids
Disk solid bowl	manual	20 – 100 000	< 1	0	1000 – 300 000	10 000	firm paste
Solids-ejecting, radial	intermittent	20 – 100 000	< 25	< 3000	1000 – 500 000	14 500	thick flowing slurry
Solids-ejecting, axial	intermittent	1000 – 150 000	< 15	< 1000	110 000 – 220 000	15 000	thick flowing slurry
Nozzle, pressurized-discharge [b]	continuous	1000 – 180 000	4 – 30	> 150 < 40 000	69 000 – 180 000	15 000	thick flowing slurry
Peripheral nozzle	continuous	300 – 500 000	2 – 30	> 3000 < 140 000	35 000 – 180 000	11 000	thin slurry
Decanter centrifuge	continuous with scroll	300 – 200 000	5 – 50	< 50 000	400 – 25 000	10 000	thick slurry to semisolid
Tubular bowl	manual	20 – 7000	< 1	0	1400 – 4500	31 000	firm paste
Multichamber bowl	manual	100 – 20 000	< 5	0		9000	firm paste
Centritech machines [c]	intermittent	5 – 100	< 1	< 15	–	100	very thin slurry
Inverted solid bowl	intermittent with scraper	10 – 6000	1 – 30	< 1000	800 – 20 000	20 000	firm paste

[a] Low-speed imperforate tubular and basket machines not included.
[b] Only for slurry of single cells (see text).
[c] Only for mammalian cells.

This type of machine was recently introduced for cleaning of crank-case gas in truck diesel engines. Oil droplets and soot particles are separated. The unit is driven by pressurized oil from the engine lube oil flow [3].

Solids-ejecting separators are the most common disk-bowl centrifuges. Automated intermittent discharge of the separated solids accumulated at the periphery allows relatively high concentrations to be obtained in the discharged solids. The rheology of the solids is very important for successful application of this type of machine. Dilatant solids such as starch cannot be discharged; the most suitable materials are noncrystalline solids of biological origin. Mixed solids can often be discharged but may require that a large amount of liquid is discharged together with the solids and/or that the residence time of the solids in the bowl is short, thus avoiding the formation of excessively hard packed solids.

Figure 4 shows a solids-ejecting bowl. The discharge takes place in about 0.1 s by lowering a hydraulically operated slide at the bottom of the bowl to release solids through ports in the bowl body.

This is illustrated in a video clip.

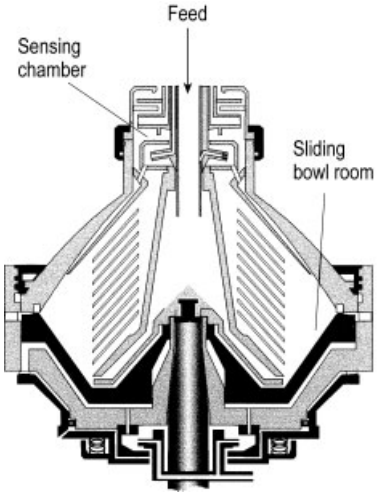

Figure 4. Radially solids-ejecting disk bowl, top feed (vertical section)

The version of bowl shown is equipped with a sensing chamber, which senses the level of solids in the bowl and triggers a discharge when the solids space is full. The solids space is designed so that solids can slide easily towards the opening. For solids with good flow properties, the angles between the solids space wall and the axis of rotation can be made smaller and hence

the outer diameter of the disks larger, increasing the separation area.

In applications where the particles in the feed material are easily broken up by the shear forces in the inlet, machines with feed from below through a hollow spindle are commonly used. In this case the bowl is filled with liquid into the center. For liquid – liquid separation, the split between the flows can be made by controlling the pressure applied on the exiting liquids. Solid-bowl machines are also available with this system.

Solids-ejecting bowls are made with diameters ranging from 180 to 1000 mm, (Figs. 5, 6, and 7). For further data, see Table 1.

Figure 5. Solids-ejecting disk-bowl centrifuge of pilot size for sterilizable contained installation (type Culturefuge 100; courtesy Alfa Laval Separation AB)
The separator itself is 1.3 m in height.

Another design of an intermittent discharge system is shown in Figure 10 where it is used as cleaning system for a nozzle machine. It is also available on its own and is suitable for discharging solids with very good flow properties, e.g., proteins and microorganisms. Solids are discharged from a number of peripheral axial channels, which are opened by lowering a slide with valve seats. The valve slide operates by allowing pressurized air to enter the volume between the bowl bottom and the slide. Models with a hydraulic discharge system are also available. Bowls with diameters of between 530 and 900 mm are currently in production. The machines have high speeds and high *g* forces (Table 1).

Figure 6. Solids-ejecting disk-bowl centrifuge for medium-scale production and explosion-protected and inerted installation (type CHPX 510; courtesy Alfa Laval Separation AB)
The height of the separator is 1.4 m. In the foreground is a liquid seal tank.

Figure 7. Solids-ejecting disk- bowl centrifuge of a large production size (type CHPX 517; courtesy Alfa Laval Separation AB)
Hydraulic capacity 40 m^3/h. The height of the separator is 1.7 m.

Nozzle Machines. When the solids flow rate is too large or the solids rheology unsuitable for the solids-ejecting machines, peripheral nozzle machines are used (Fig. 8). A possible alternative to nozzle machines are decanters, but the separation area of these is more than an order of magnitude smaller, and they are therefore used when the separability of the solids is better. The number of nozzles increases with increasing bowl size and ranges from 4 to 20. The nozzle diameter ranges from 0.5 to 3.5 mm.

The approximate flow Q_N through nozzles in nozzle machines can be calculated from the formula

$$Q_N \approx k\, a_N\, r_N\, \omega\, z \;\; [\mathrm{m}^3/\mathrm{s}] \tag{17}$$

where a_N is the nozzle area, r_N the distance of nozzle from center of rotation, ω the bowl speed in rad/s, z the number of nozzles, and k is a proportionality constant ($k \approx 0.83$) (all quantities in SI units). The nozzle flow is proportional both to the radius and to the bowl speed.

Some bowls are equipped with a central feed tube for recirculated solids. These solids are introduced in front of the nozzles, bypassing the disk stack, thereby avoiding disturbances to the separation process. The recycle stream is used to increase the solids concentration when insufficient solids are present in the feed to fill the smallest possible nozzles. Small nozzles have a larger tendency to be clogged by oversize particles in the feed. Many nozzle machines are used without recirculation of solids, and the same bowl can be used but without the recycle components. In a third variant, the concentrate tubes can be used to feed wash water to the nozzles from the central chamber, which is supplied with water through a hollow bowl spindle.

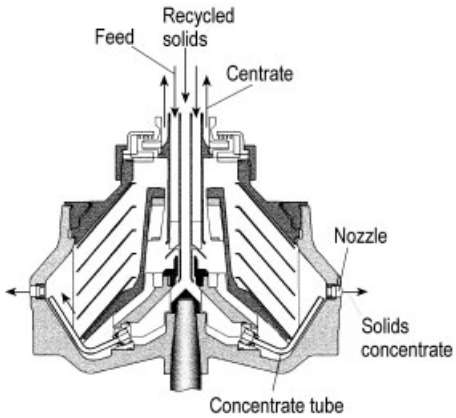

Figure 8. Nozzle bowl with peripheral discharge nozzles and recirculation of solids (vertical section)

Figure 9. A nozzle machine type TX 320 with peripheral nozzles installed in a starch factory (courtesy Alfa Laval Separation AB)
Hydraulic capacity 200 m^3/h. The height of the machine is 2.2 m.

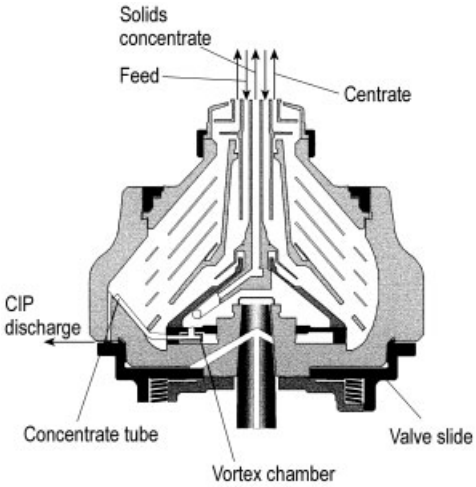

Figure 10. Nozzle bowl with pressurized nozzle discharge and slide for Cleaning-in-Place (CIP) (vertical section)

Peripheral nozzle sedimenting centrifuges are suitable for the largest flow rates (see Table 1), and the bowls can have a diameter of up to ca. 1000 mm. A large nozzle machine of this type is shown in Figure 9.

With pseudoplastic solids having an angle of repose of 0° (e.g., yeasts and bacteria), the solids concentrate is taken from the periphery in a number of tubes (Fig. 10) into a chamber in which it is taken up by a nonrotating paring tube, which pumps it up and out through a central pipe. In front of the nozzles, vortex chambers can be mounted. The concentrate enters these tangentially and creates a whirl which acts as a resistance to flow over the chamber. With high-viscosity concentrate, the speed of the whirl is small, the flow resistance is low, and flow through the system is high. At low viscosity, the flow is higher. The viscosity effect gives rise to a self-regulatory function, so that with variable solids flow (within limits) into the machine, the resulting concentration in the solids concentrate is constant. The risk of clogging the machine is reduced. In machines without vortex chambers, the concentrate tubes can be replaced by channels in the bowl body. These channels start at the periphery of the bowl, are directed towards the center of rotation, and end with nozzles at the bottom of the bowl, spraying out the concentrate under the bowl.

Figure 11. Nozzle machine with pressurized solids discharge for yeast and bacteria separation. Capacity for *E. coli* bacteria about 3000 L/h (type BTUX 510; courtesy Alfa Laval Separation AB)
The height of the separator is 1.8 m. Sterilizable and contained installation.

Bowl diameters between 528 and 900 mm are manufactured. A medium-sized machine of this type is shown in Figure 11.

3.3. Chamber and Multichamber Bowls

General. A chamber bowl has the shape of a short cylinder with an axial length approximately equal to its diameter. This geometry is used in the most diverse fields, from low-speed versions in mechanical workshops to ultracentrifuges for separation of individual proteins.

Imperforate-Basket Machines. The principle of a simple basket centrifuge is shown in Figure 12. When the basket is full of solids, it must be emptied manually. A high-speed titanium bowl with semi-automatic intermittent solids removal is shown in Figure 13 [18]. After a suitable time, the bowl speed is decreased, and an asymmetrically positioned scraper is engaged and removes the solids. Bottom-fed bowls with similar solids-removal mechanisms are also available. Accelerating vanes are not possible with these mechanisms. Basket-type bowls are also used for liquid – liquid separations, e.g., as centrifugal settlers in countercurrent multistage extraction systems. The mixer often corotates with the bowl on the same spindle.

The Liquid – Liquid Extractor. One of the most widely used multistage extractors can be regarded as a chamber centrifuge with bowl inserts [19]. It is essentially a sieve-plate column wrapped around a shaft. It has a horizontal axis with bearings on both sides of the rotor. It is a hermetic machine in which the feed and the solvent enter the rotor through hollow shafts from opposite sides of the rotor. Feed flows of up to 60 m^3/h can be handled.

Preparative continuous-flow centrifuges can be classified as chamber bowl machines with a completely filled bowl. This is accomplished by the inclusion in the design of mechanical face seals between the rotor and the inlet/outlet. In the bowl is an insert, the core, with wings (septae) that aid in accelerating the liquid. In the wings, channels are drilled for feed and centrate liquids (Fig. 14).

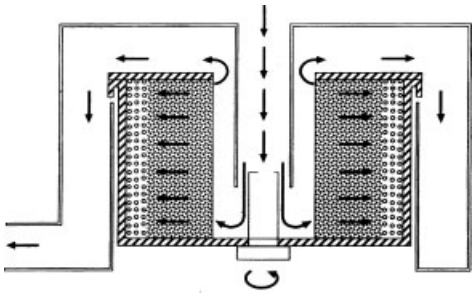

Figure 12. Bowl principle of an imperforate basket machine (vertical section)

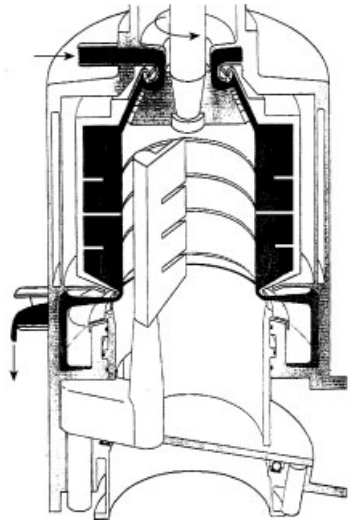

Figure 13. Imporforate basket machine with intermittent solids removal (courtesy of CARR Separations, Inc) Vertical section.

The bowls are made of titanium and have volumes of up to 1.7 L; Z is up to 102 000. The inner bowl diameter is up to 190 mm; the sedimentation path is 12 – 59 mm. The machines are used for the sedimentation of large viruses, bacteria, or cell constituents directly to the peripheral bowl wall, or as zonal centrifuges (see Sections 4.2 and 4.3). In the latter machines, the particles of interest are either collected within a density gradient, or on to a cushion of high-density fluid (e.g., sucrose solution) that protects sensitive particles from being destroyed at the bowl wall (see legend to Figure 14) [20]. Especially the machines with the highest g forces must be refrigerated or run under vacuum to avoid heating.

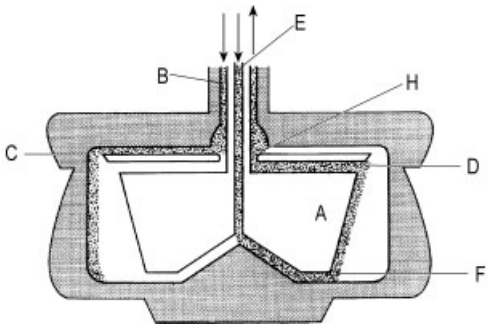

Figure 14. Continuous-flow operation of the CF-32-Ti and JCF-Z (standard pellet core) zonal rotors (from [20], courtesy Oxford University Press)
(A) rotor core and septa. The annular channel (B) forms the common external path to the edge of the rotor (C) and the top of the core (D). Gradient is fed to the wall of the rotor via (B) and (C). Sample passes down the central channel (E), and up over the surface of the core between (F) and (D); the clarified supernatant exits at (G). To unload the rotor an air lock at (H) ensures the passage of dense sucrose to the wall from (B) to (C) and the effluent is collected at (E).

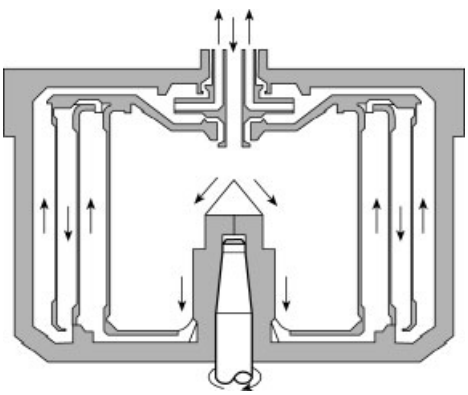

Figure 15. Multichamber bowl (vertical section)

Multichamber Bowl. In this machine, the process liquid flows in the annuli between concentric cylinders (Fig. 15). Having a relatively low separation capability, the machine is used primarily for recovery of valuable solids in dry form and in larger volumes than with the tubular bowl. The maximum bowl diameter is ca. 500 mm, and the solids volume is up to 50 L.

3.4. Tubular Bowl Centrifuges

General. The rotor consists of a long cylinder, into which liquid is pumped (Fig. 16). The tubular bowl geometry is used in low-speed, high-speed, and ultrahigh-speed applications. The high *L*/*D* ratio allows bowl volume and retention time to be increased without sacrificing bowl speed, which can be high if the diameter is kept small.

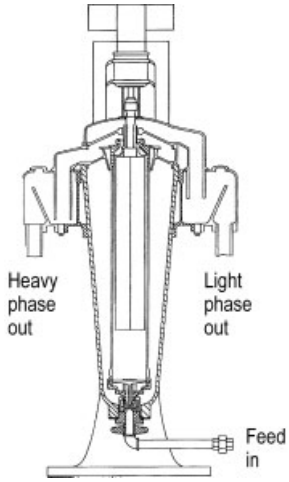

Figure 16. Tubular-bowl centrifuge (vertical section)

High-Speed Applications. The design shown in Figure 16 is designed for separation of two liquid phases. To remove accumulated heavy solids, the machine must be stopped and dismantled. The length to diameter ratio is usually up to 7 : 1. The small outer diameters make high *Z* values possible (see Table 1). In the high-speed machines, the bowl diameter is up to 130 mm, and the *L*/*D* ratio up to 7. The solids collecting volume is up to 5 L.

Ultrahigh-Speed Applications. For large-volume production the continuous preparative centrifuge shown in Figure 14 has too small a bowl volume. Therefore zonal centrifuges with a large *L*/*D* ratio have been developed. The feed and outlet lines are in principle arranged as in Figure 14. The rotor diameter can be about 130 mm and the rotor volume up to 7 L [5, 21].

Low-Speed Applications. In a design for liquid – liquid separation in multistage extraction plants [22], bowl diameters up to 500 mm are available. The bowl speed is very low, so that *Z* is well below 1000. The mixing of the two phases takes place in the annulus between the rotor and the housing. A similar design has been

used for several years for nuclear reprocessing [23].

Gas Centrifuges. For the enrichment of ^{235}U a number of different techniques have been devised. One is to make UF_6 gas and exploit the density difference between $^{235}UF_6$ and $^{238}UF_6$. A small enrichment is achieved in each separation step, but thousands of units are needed, coupled in series [24]. A modern gas centrifuge may have a diameter of 500 mm and a length of 5000 mm and develop a Z of 147 000 [13]. Development of more efficient rotors is continuing, but details are not in the public domain.

3.5. Decanter Centrifuges

General. The bowl of a decanter centrifuge generally consists of a cylindrical and a conical part. A coaxially mounted internal rotor with helical flights allows for axial transport of the solids by rotation of the conveyor relative to the bowl. The bowl and the conveyor are connected through a gear box which is used to control the relative movement of the conveyor.

Depending on the direction of rotation and whether the conveyor flight is right- or left-handed, the conveyor must rotate faster or slower than the bowl to ensure solids transportation in the correct axial direction. These two designs are known as trailing and leading conveyors, and they result in significantly different flow patterns in the decanter centrifuge (see Section 5.3).

The rotating parts are mounted on a stationary frame, which often carries the drive motor and a back drive system that controls the rotational speed of the sunwheel shaft of the gear box. The frame is connected to the ground through vibration isolators. Decanters are used for liquid – solid and liquid – liquid – solid separation. In the latter case the interface position is given by Equation (10).

The liquid outlet is usually an open outlet with adjustable weir plates, but a stationary paring disk can also be employed. For three-phase operation (liquid – liquid – solid) the choice is between two open outlets or an open outlet combined with a paring disk. In some designs the liquid is picked up in the middle part of the decanter centrifuge through return tubes mounted on the conveyor. In such designs, the inlet is located at the cylindrical end of the conveyor; the liquid and the solids travel cocurrently. In the more popular design with the inlet located near the intersection of the cylindrical and the conical parts of the bowl, the liquid and the solids travel countercurrently.

Horizontal decanters. Horizontal decanters are by far the most common type of decanter centrifuge. The process flow through the decanter is shown in Figure 17. The suspension is pumped through a stationary feed tube into an inlet in the conveyor body where it is accelerated and fed out to a ring-shaped pond of liquid. The liquid flows through the helical channel towards the large end hub, while the solids separate out and settle on the inside of the bowl wall. The clarified liquid leaves the decanter centrifuge over adjustable weir plates, and the relative rotation of the conveyor transports the solids towards the conical end. The solids are conveyed inwards and lifted out of the pond before they leave the centrifuge through openings in the conical part.

The mode of operation is shown in the following video clip.

The conveyor shown in Figure 17 has solid flights which ensure fully helical flow in the pond. Depending on the application the flights often have openings that allow for some axial flow or, in some designs, even entirely axial flow.

In a decanter centrifuge, the radial distance from the center of rotation to the inner liquid surface is large compared to the distance in a disk-bowl centrifuge of the same outer diameter. This is a consequence of the need for space inside of the liquid surface for transporting a large amount of solids without the risk of liquid escaping through the solids discharge ports. The early decanters were mainly conical with a short cylindrical part, if any. The solids handling capacity was high compared to the sedimentation capacity. The main development focus up until the late 1980s was to increase the sedimentation capacity while maintaining solids-handling capacity. The feed suspension for decanters often contains large particles; hence, a disk stack cannot be generally employed. The large radial distance from the center of rotation to the liquid surface further reduces the sedimentation capacity due to high turbulence level. Instead, the sedimentation capacity can be increased by increasing the length of the decanter (see Section 5.3). A more detailed discussion of the sedimentation

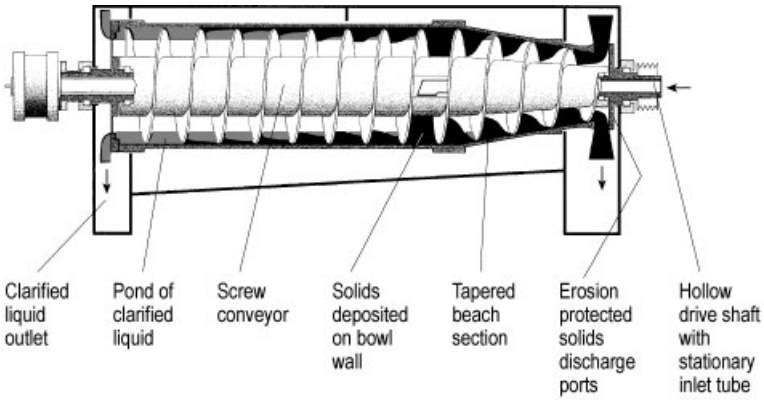

Figure 17. Process flow through a decanter centrifuge

performance is given in [25]. A discussion of mechanical and rotor dynamical limitations for long decanters is found in [26]. Since the mid- to late 1980s the dryness or the quality of the solids has governed the development of new decanter designs.

The best quality of the solids usually means the driest solids, but in a few applications, e.g., the thickening of activated sludge, the cake should still be flowable, and the desired solids concentration can be as low as 6 wt %. Such fluidlike solids are difficult to convey by differential movement alone. Since the late 1970s decanters for this application have been equipped with a baffle disk.

Figure 18. Baffle disk designs
Top: Baffle disk; Middle: Baffle plate; Bottom: Conveying baffle.

Figure 18 shows three types of baffle disks: disk, plate, and conveying baffles. The disk and plate baffles have been known since the late 1970s; the conveying baffle was introduced in the mid-1990s.

For rotation of Figure 18 see the following videoclip.

The baffle disk restricts the passage to the solids outlet such that the solids must travel through a relatively small clearance between the outside of the baffle disk and the inside of the bowl wall. The difference in hydrostatic pressure between the liquid and the solids side of the baffle disk transports the solids through the clearance and out of the bowl. When the solids are removed by pressure instead of pure conveyance, the long conical part is no longer needed, and it is often replaced by a short conical part (steep cone, see Fig. 19) with a half-cone angle of 12 – 20°, as opposed to 6 – 12° for other decanters.

The concentration of compressible solids has also greatly influenced the development of decanters. The largest application with compressible solids is the dewatering of municipal sludges.

When solids consist of discrete particles, the dryness can normally be increased by improving the drainage conditions after the solids leave the pond. However, with compressible solids no significant drainage takes place here. Compressible solids are better suited to compaction below the liquid surface. To maximize the load on the solids the decanter must be operated with a thick layer of solids in the cylindrical part. The baffle disks used for thickening are also applicable to dewatering, where they maintain a thick layer

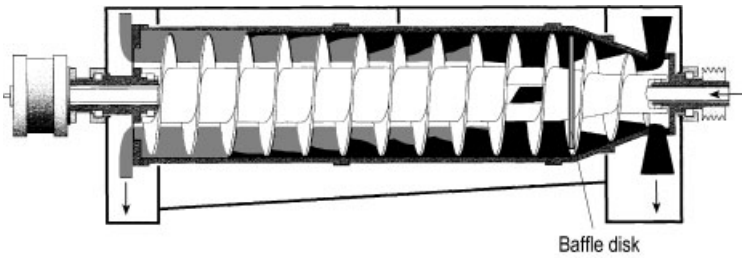

Figure 19. Baffle disk operation with steep cone

of solids in the cylindrical part. With a well-controlled flow of solids across the baffle disk, the volume of solids in the decanter can be further increased to give the maximum retention time without disturbing sedimentation.

The *L*/*D* ratio of the bowl gives a good indication of the relation between solids handling capacity and sedimentation capacity of a decanter. Solids handling capacity depends mainly on the bowl diameter, while the sedimentation capacity depends on both the diameter and the length. For decanters with a trailing conveyor the upper limit for *L*/*D* is around 4.2 at high speed. Lowering the speed or adopting special design solutions has made a few models with *L*/*D* = 5.2 available. For leading conveyor designs the maximum *L*/*D* is generally between 3 and 4.

Horizontal decanters are available in diameters ranging from 150 to 1500 mm, covering flow ranges from 0.3 to 200 m^3/h.

Vertical Decanter Centrifuge. Decanters are also designed with a vertical axis of rotation. This is generally more expensive and offers no immediate advantages over horizontal decanters of the same size. Vertical decanters are often installed in the chemical industry when special requirements are involved:

1) Vaportight or pressuretight casings
2) Operating temperatures above 100 °C or below − 25 °C
3) Very corrosive or toxic operating conditions

Due to the differences in vibration isolators, bearings, and especially seals, the vertical design has advantages under these special conditions. Figure 20 shows the main design differences between horizontal and vertical decanters. Vertical decanters are available in diameters from 180 mm to 750 mm.

Disk Decanter Centrifuge. The newest addition to the decanter family is a vertical decanter equipped with a disk stack [27]. The rationale is to combine the best characteristics of a decanter (dry solids) and of a disk bowl machine (good clarification).

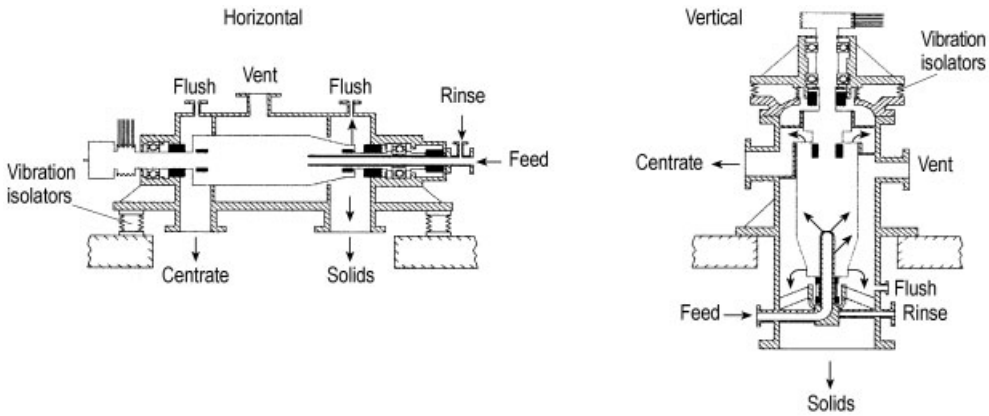

Figure 20. Main design differences between a horizontal and a vertical decanter

3.6. Other Bowl Geometries

A type of field flow sedimentation takes place in the plastic bladder shown in Figure 21. The machine was developed for separation of shear-sensitive mammalian cells, and Z is rather low, about 100 [28]. The bladder is inserted in a rotor, and no seals are needed between the stationary parts and the rotor.

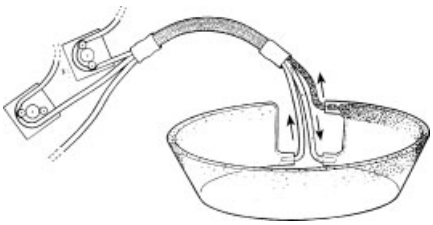

Figure 21. The separation cell of a Sorvall Centritech cell centrifuge (courtesy Sorvall Inc.)

The feed liquid enters at the bottom of the insert and the main liquid outflow is at the opposite end of the bottom of the insert. The heavy liquid phase (cell concentrate) is withdrawn intermittently through the top outlet.

4. Batch Centrifuges

4.1. General

In batch centrifuges a sample of the liquid to be separated is added to the centrifuge container either at standstill or in rotation, and the whole volume remains in the container until the centrifugation period is over. There are two general categories: bucket or bottle centrifuges and batch zonal centrifuges.

4.2. Bottle Centrifuges

Bottle centrifuges are classified as shown in Table 2.

Table 2. Classification of bottle centrifuges (adapted from [21])

	Low-speed	High-speed	Ultracentrifuge
Speed range, 10^3 rpm	2 – 6	18 – 30	35 – 120
Maximum Z	8000	100 000	700 000
Refrigerator	some	yes	yes
Acceleration/Braking	some	variable	variable

The ultracentrifuges can be divided into analytical and preparative ultracentrifuges. The preparative centrifuges, which can belong to all categories in Table 2 are used in research and in the production of biological materials, especially vaccines. The low-speed centrifuge is a valuable analytical tool for characterization of process liquids in the procedures for selecting large-scale centrifuges.

The analytical centrifuge is a sophisticated ultracentrifuge that is used to study optically the sedimentation of small samples (< 1 mL) to determine the molecular mass of macromolecules (→ Plastics, Analysis, Chap. 5.1.3). This method was of major importance in biochemical research, but the emergence of other techniques has restricted its application. More information is given in [1, 5], and [21].

Preparative centrifuge rotors can be classified as swing-out (Fig. 22), fixed-angle (Fig. 23) and vertical (tube axis parallel to the rotor axis). Table 3 summarizes some data for commercially available high-speed and ultracentrifuges [29]. A large number of tubes means a low tube volume, and a large total tube means a lower g force.

Bucket centrifuges can be used for a number of separation methods:

1) Differential centrifugation (pelleting)
2) Density-gradient centrifugation
   (a) rate-zonal (sedimentation velocity) method
   (b) isopycnic (sedimentation equilibrium) method

In pelleting the tubes are filled with a uniform sample. During centrifugation, particles sediment to the bottom of the tube and form a layer or a plug of sediment (see Section 2.1).

Pelleting is too crude a method for many separations of biological materials. Therefore density-gradient separation techniques have been developed in which sedimentation takes place in a medium whose density increases with increasing distance from the center of rotation. This permits separation of several or all components in a sample and also analytical measurements of, for example, sedimentation rate, density, and molecular mass. The sample volumes that can be handled with this technique are, however, an order of magnitude smaller than in pelleting.

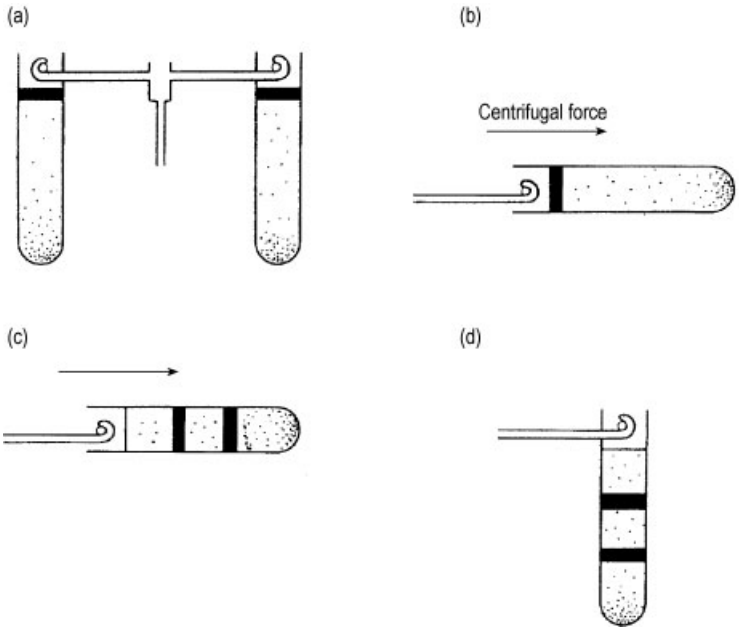

Figure 22. Separation in swing-out rotors
The tubes are filled, loaded into each bucket and attached to the central body (yoke) of the rotor; at rest the buckets of the rotor hang vertically (a). As the rotor begins to move the buckets move out so that they are perpendicular to the axis of rotation (b). During centrifugation, the particles sediment down the tube (c). When the rotor comes to a stop (d) the buckets return to a vertical position; there is no reorientation of the liquid in the tubes (from [21], courtesy Oxford University Press)

Table 3. Characteristics of high-speed and ultracentrifuges (adapted from [29])

Rotor type	Number of tubes	Tube volume, mL	Rotor speed, 10^3 rpm	$Z \cdot 10^{-3}$
Fixed angle, small volume	6 – 100	0.2 – 13.5	25 – 120	70 – 700
Fixed angle, medium volume	8 – 12	36 – 50	28 – 70	90 – 500
Fixed angle, large volume	6 – 10	90 – 250	16 – 45	40 – 250
Swing-out, small volume	3 – 6	2 – 5	30 – 65	120 – 500
Swing-out, medium volume	6	13 – 20	28 – 41	120 – 300
Swing-out, large volume	6	20 – 40	25 – 30	90 – 150
Vertical	8 – 12	2 – 39	50 – 100	240 – 645

In the rate-zonal method, the sample is layered on top of the density-gradient liquid. During centrifugation each particle type sediments down the tube according to its sedimentation rate. The density of all sample particles of interest must be larger than the density of the liquid at any specific point in the tube during the whole centrifugation procedure.

In the isopycnic method each particle sediments to the point where its density is equal to that of the fluid.

4.3. Batch Zonal Centrifuges

In density gradient centrifugation in bottle centrifuges, the amount of particles that can be collected is limited by the cross-sectional area of the tube. For larger quantities of sample, zonal rotors of cylindrical or bowl shape have been developed that can handle 50 – 100 times the volume in bottle centrifuges.

5. Fluid Dynamics in Centrifuges

The mathematical equations describing the flow in centrifuge bowls are very complicated, and therefore it is not easy to solve them. The presence of solids complicates matters even more. For mathematical treatments, see [5, 10].

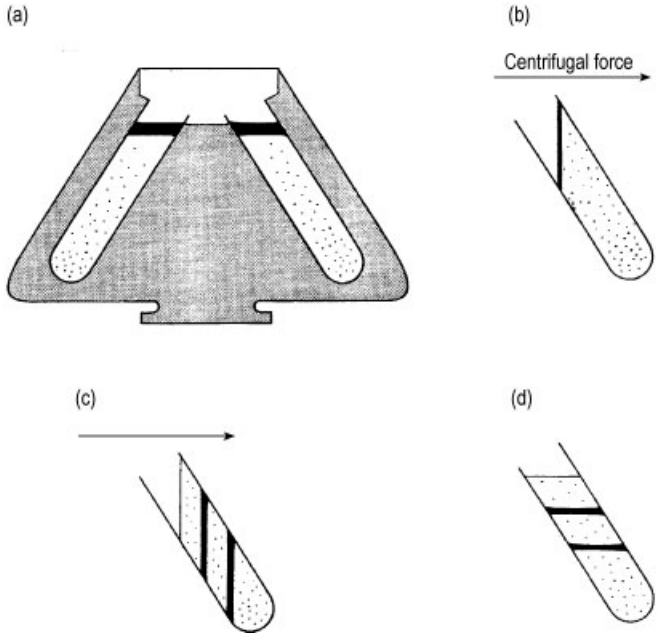

Figure 23. Separation in fixed-angle rotors
The tubes are filled and placed in the rotor pockets (a); tubes for an ultracentrifuge are filled completely and sealed, unsealed tubes must not be overfilled. As the rotor begins to move, the liquid in the tube reorientates during acceleration (b) and the particles sediment down the tube (c). As the rotor comes to a stop (d) the liquid in the tubes reorientates as gravity becomes stronger than the centrifugal force (from [21], courtesy Oxford University Press).

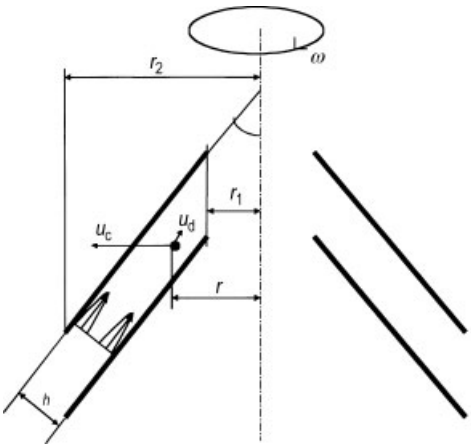

Figure 24. Flow profile between conical discs at high λ values. Particle separation.

5.1. Chamber and Tubular Bowl Centrifuges

The fluid dynamics is complicated even in the simplest of bowl geometries. Sokolov [2] compared actual separation performance with those predicted by a number of velocity distributions. He found that the best fit was obtained with a plug-flow model, but the separation was overestimated. The phenomena that make the simplified flow models inaccurate are end effects, vortices between the wings of a wing insert, velocity lag, and nonideal feed accelerators.

5.2. Disk-Bowl Centrifuges

The disks in the bowl split the flow into thin layers. Usually the flow between the disks is laminar, which facilitates particle settling. The critical Reynolds number for transition to turbulent flow increases with decreasing disk spacing and increasing bowl speed. This is also beneficial since a short disk spacing makes the settling distance short, and the high bowl speed increases the settling velocity. Figure 24 illustrates the situation. The liquid is driven towards the center by a pressure gradient. The particle, in this case denser than the liquid, is considered to be separated if it reaches the underside of the upper disk. The particle movement is influenced by centrifugal force and by the liquid drag. The flow pat-

tern of the liquid is determined by the interplay of inertial, friction, and Coriolis forces. The flow profile is determined by a dimensionless number λ, defined as

$$\lambda = t \cdot \sqrt{\frac{\omega \sin \alpha}{u}} \tag{18}$$

When λ is small the velocity distribution is parabolic, but for values > 5, normally encountered in industrial centrifuges, the velocity profile breaks up in two so-called Ekman layers, in which the net transport of liquid towards the center takes place. Since separated particles (heavy phase) are collected at the underside of the disk, they are subjected to a strong shear force from the Ekman layer that drags the particle towards the liquid outlet. The velocity in these layers increases with increasing λ, that is, with increasing bowl speed. Therefore a higher bowl speed cannot always be fully utilized because of the counteracting shear force.

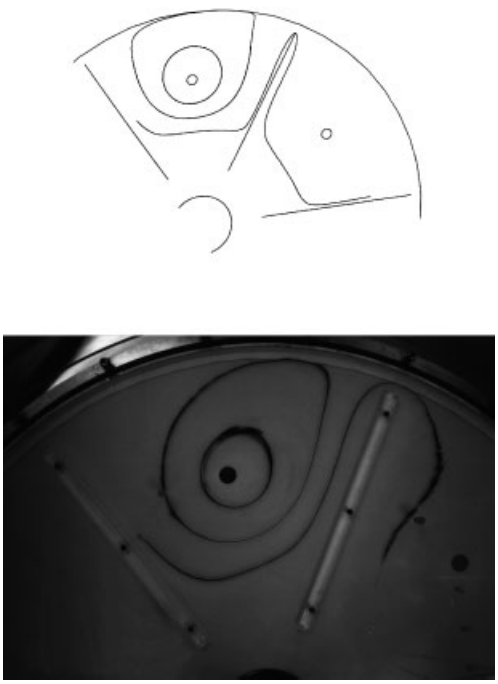

Figure 25. Comparison of calculated (a) and actual flow patterns (b) in a disk bowl [30]

In between the Ekman layers is the geostrophic layer. In Figure 25 its streamlines have been made visible by dye injection in the geostrophic layer and can be compared with the calculated streamlines in the gap between two discs with radial spacers and a distribution hole through which the liquid enters [30]. Particle separation has also been studied in the same geometry [31]. It has been found that the drag force on the particles collected at the underside of the disk depends on the position on the disk, and that some particles are easily separated, whereas others are forced towards the center. The drag force can be computed [32] and the calculations of separation efficiency including the drag force verify these findings.

Many bowls have an open chamber in the center in which the nonrotating feed liquid is received. In many cases the turbulence in the feed zone breaks up shear-sensitive particles, making separation more difficult. For many applications it is therefore advantageous to use a machine with feed from below. In other applications a special inlet geometry (e.g., a disk inlet) dissipates the energy over a larger volume and decreases the shear forces [33]. It can also reduce air entrainment and foaming.

5.3. Decanter Centrifuges

The liquid flow patterns in a decanter centrifuge have been studied by using a transparent bowl operating at reduced speed. One area of particular interest is the difference between trailing and leading conveyor design. Operation of full-scale decanters of the two types have often revealed some unexplained differences. Figure 26 shows photographs of the flow patterns in the transparent bowl for the two design types. In both cases the decanter operates with a constant feed flow of water, and the inlet position is close to the intersection between the cylindrical and the conical part. At $t = 0$, a pulse of blue color is added to the feed stream and the position of the colored liquid is photographed after 5, 15, 30, and 60 s. The findings can be summarized as follows:

In the *trailing conveyor,* the feed travels rapidly away from the feed position in a thin surface layer in the direction of the liquid outlet. As the liquid flows along the helical path the friction against the flights gradually changes the velocity profile into an increasingly accelerated stream. As the acceleration increases, the depth of the surface layer also increases, and the velocity towards the outlet drops significantly.

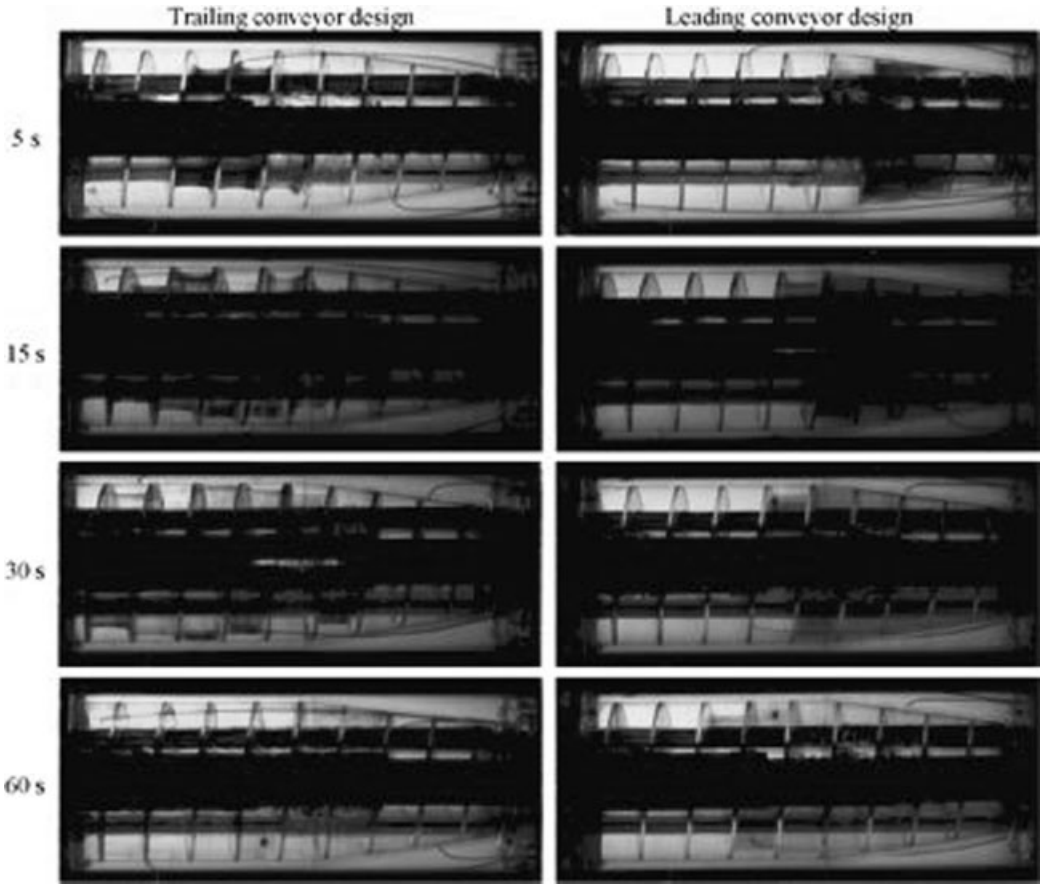

Figure 26. Flow patterns in decanter centrifuges
A pulse of blue color is added to the constant flow of water at $t = 0$. Photos are taken at 5, 15, 30, and 60 s with stroboscopic light from behind. With the trailing conveyor design the flow starts as a thin surface layer flow at the feeding position. As the liquid travels towards the outlet, the flow profile develops into a less turbulent profile that is better suited for sedimentation. A substantial part of the liquid is recirculated back for remixing at the inlet. In the leading conveyor design the liquid travels in the wrong direction at the surface at the entrance in the pond. After being brought up to full rotational speed, it moves in the right direction in a layer below the surface and flows almost as a plug flow in the cylindrical part.

The existence of a surface layer is well known from the literature, but in many cases depth of the layer is assumed to be constant. This leads to a poor description of the sedimentation capacity of long decanters as it is the change in velocity profile that ensures constantly improved sedimentation conditions with increasing distance from the inlet. The variation in velocity profile also includes a return flow, in which a significant amount of liquid flows back to the inlet, where it re-entrains the incoming feed stream. The return flow to some extent protects the solids on the bowl wall from being dragged along in direction of the main flow. (No solids are present in the test shown in Figure 26.) Due to the improved sedimentation capacity of long decanters, the L/D is often chosen to be as high as possible for this design type. The limiting factors are the solids handling capacity and rotor-dynamics considerations.

In the *leading conveyor,* the surface layer starts out in the wrong direction. It has to come back again in a subsurface layer and, after passing the inlet position, where some of it is re-entrained, it flows almost like a plug flow with constant velocity towards the liquid outlet. The plug flow velocity profile is ideal for sedimentation, and this design type has quite a good sedimentation capacity even for short machines (low L/D). The limitation of this design is that the

full feed stream must pass the inlet twice, which reduces the space available for the transport of solids. Furthermore the direction of the solids is opposite to the returning feed flow, and this limits the amount of solids that can be transported past the critical area. As a result decanters with a leading conveyor have smaller values of *L*/*D*.

The *presence of solids* will, of course, also influence the flow patterns, especially when they take up the majority of the pond volume, and this further complicates the fluid dynamics of the decanter centrifuge. Such conditions cannot be studied in the transparent decanter shown in Figure 26, but full-scale tests with low solids load confirm the general impression from the observed flow patterns [25]. As the flow pattern inside the decanter depends on the design type (leading or trailing conveyor) and on *L*/*D*, one has to be very careful when scaling from one decanter to another. Generally it is recommended only to scale between the same design type and similar *L*/*D* values unless the scaling is supported by experience with the particular application.

6. Centrifuge Technology

6.1. Rotor Dynamics

All rotating bodies vibrate at certain speeds. At these critical speeds the rotational frequency of the assembly is equal to one of its natural frequencies. The operating speed must therefore be selected so that it differs from these frequencies and their harmonics. At a critical speed any unbalance load will cause heavy vibrations and high stress to the bearings and the supporting system. If the rotor at a critical speed hits nonrotating parts of the assembly the deflections will suddenly be even greater and may lead to even more serious damage. Modern centrifuges are therefore designed with sufficient clearances between rotating and nonrotating parts that contacts are very unlikely.

Centrifuge rotors are of two kinds: those operating with fixed bearings at subcritical speed and those mounted with flexibly mounted bearings for operation at supercritical speed. The latter category includes disk-bowl (Figs. 2 – 11), multichamber-bowl (Fig. 15) and high-speed tubular-bowl (Fig. 16) machines. Lately, decanters with supercritical operation have been built [26]. Otherwise, all decanters are designed for subcritical operation.

In the supercritical machines, the deflections experienced when passing the critical speeds at starting and stopping can be dampened by mounting the bearings on springs or rubber elements.

6.2. Materials Selection

Bowls. In selecting a material for a centrifuge bowl, a number of conflicting demands have to be met. The material must have high yield stress, high tolerance to fatigue, high corrosion resistance, and low cost.

Although exceptions exist, the centrifuge designer strives for the highest possible *g* number combined with a bowl volume large enough to provide an optimal residence time or solids space. For steels in general, the higher the corrosion resistance is, the lower is the strength. The tendency today is to increase yield stress in materials with the required corrosion resistance, because this is in many cases the cheapest way to improve the cost/performance ratio.

In disk-bowl and decanter centrifuges, Duplex steels are very common [34]. In applications with moderate corrosiveness, a DIN 1.4462 steel (in USA S31803, in Sweden 2377) is often used. Its yield strength $R_{p0.2}$ is 450 MPa. Other Duplex steels are under development [35]. For applications of low corrosiveness and low chloride ion content, DIN 1.4418 steel (in Sweden 2387) with $R_{p0.2}$ = 750 MPa is often employed. Unalloyed steels are used in some decanters and low-cost basket centrifuges. Hastelloy is used for extreme conditions in the chemical industry; the yield strength is low compared to the other materials mentioned above. For special purposes, aluminum alloys can also be used. Alloyed titanium (typical $R_{p0.2}$ = 830 MPa) is the bowl material in batch and continuous-flow zonal centrifuges and, together with aluminum, in medium- and high-speed bucket centrifuges. The bowl of the basket machine in Figure 13 is made of titanium for use in the biotechnology and fine-chemicals industries.

The stresses in the bowl material are the result of the weight of the material itself and the bowl contents. The self-stress accounts for 50 – 65 %

of the tangential stress; the remainder is due to the weight of the bowl contents (density 1000 – 1500 kg/m^3). It can be shown [13] that the maximum possible peripheral speed $\omega_{\max} r_{\text{outer}}$ is approximately proportional to the square root of the yield strength. Therefore $(\omega_{\max} r_{\text{outer}})^2 = Z_{\max} r_{\text{outer}} = \text{const.}$, and $Z_{\max}$ is approximately inversely proportional to the outer bowl radius. This is valid for geometrically similar machines.

The maximum peripheral speed of a solids-ejecting bowl (Fig. 4) made of DIN 1.4462 steel is about 170 m/s. For other geometries with this material (e.g., Figs. 2 and 10) it is somewhat higher. Zonal rotors made of titanium have maximum peripheral speeds of the order of 300 m/s. Gas centrifuges can run even faster, partly on account of a lower density of the bowl contents; a peripheral speed of 600 m/s has been reported [13].

The axial stresses on the bowl are one order of magnitude smaller than the tangential stresses.

The fatigue tolerance of the material must be taken into account in design. Modern solids-ejecting bowls are designed such that no restrictions on number of discharges need to be imposed. Some zonal centrifuges record each cycle of use because the bowl lifetime is determined by fatigue.

Bowl parts subject to less stress are made from stainless steels (DIN 1.4401 or 1.4404; U.S. equivalents 316 and 316L), from mild steel or, in the absence of alkaline or other corrosive liquids, bronze. Disks for some small disk-bowl machines are made of plastics.

Frames. Most frames of industrial machines are made of cast iron. They are equipped with vibration-dampening rubber feet, which are bolted to the floor.

Covers. The material for the covers of industrial high-speed machines is most often stainless steel (DIN 1.4401 or 1.4404; U.S. equivalents 316 and 316L). Also carbon steel, plastics, and, in the absence of alkaline liquids, silumin are commonly used.

Solids Receptacles. In nonfood and mildly corrosive applications, painted cast iron is used. For food and corrosive applications stainless steel is common. When the solids are erosive, replaceable solids deflectors with hard surfacing or polyurethane linings are mounted.

Nozzles. In nozzle machines the discharged solids are often erosive. Therefore the nozzles have inserts of tungsten carbide, sometimes sapphire. For nonerosive solids, stainless steels are used.

Elastomers and Plastics. Nitrile, butyl, EPDM, polyurethane and fluoro rubber are used in O-rings and gaskets. For food applications, FDA and BGA approved materials may be specified. A large number of plastics materials are used. Polyamide, PTFE, and polyether ether ketone are some examples for large industrial machines. Polycarbonate and polysulfone are often used for the tubes and bottles in bucket centrifuges.

7. Centrifuge Selection

7.1. Introduction

The choice of type and size of centrifuge is a combined theoretical and practical exercise. In most cases the final choice for a new centrifuge application cannot be made without testing the type of centrifuge preliminary selected at between 25 and 100 % of the final plant flow rate. The reasons for this are at least two. One is the elimination of the technical risk. The other is so far most common in the industry producing pharmaceuticals. In pre-approval inspection, the inspecting and licensing authorities are concerned with the comparability between batches of product from the development phase and from the production phase. Since the validation procedures are costly and time-consuming, it is very important to already select at least the right type of centrifuge at pilot scale.

A combination of different types of centrifuge in series is often the most cost-effective solution for a given problem. Also a filter in series can be a valuable complement to a centrifuge.

Alternative process and plant designs should be discussed before the centrifuge type and size are finally selected.

7.2. Main Process Parameters [40]

The *feed flow rate* is often determined by the available or required time to process one batch.

The *separability* can be estimated on a bench scale by test tube centrifugation (spin test, see Section 7.3). The *flow of solids* is the feed flow rate multiplied by the percentage by volume of suspended solids. It is this solids flow rather than the suspended solids content alone that influences the choice of machine type. Different machine types differ in their ability to produce high *concentrations of separated solids* (see Table 1). The *required degree of separation* has a strong influence on the possible flow rate in a given machine. The wider the particle size distribution, the more the flow rate must be reduced to increase the degree of separation degree from, e.g., 90 to 99 %. *Solids rheology*, i.e., the reaction of the solids to pressure and shear forces, is very important for the choice of machine type. Pseudoplastic, i.e., shear-thinning, solids will flow out of the machine type shown in Fig. 10 at high solids contents. Shear-thinning solids may have an angle of repose of close to 0° (see Section 3.2). Dilatant solids can only be discharged in concentrated form without manual handling from peripheral-nozzle machines (Fig. 8) and decanters (Fig. 17). Because of the content of a valuable dissolved substance in the entrained liquid, *washing of solids* may be necessary. Another reason for washing solids is to remove impurities. This is usually done by resuspension and further separation steps.

7.3. Test Tube Centrifugation (Spin Test)

The most important task for a spin test is the *determination of separability*, i.e., the settling velocity. In practice, it is difficult to calculate, since often the information needed for the Stokes Equation is incomplete. The spin test is done by spinning a sample of the feed stream for different periods of time and/or at different g numbers. For each spin the supernatant is analyzed with regard to the process parameter of interest, e.g., residual solids content, turbidity, cell count, or content of heavy liquid. Thereafter the spin time that gives acceptable supernatant quality is recorded, and a preliminary Σ value is calculated from Equation (19).

The swing-out test tube (bottle) centrifuge is shown schematically in Figure 1. The shaded volume with height h is the volume taken for sampling. Combining Equations (9) and (13) and rearranging gives a preliminary value of $Q_{\mathrm{theor}}/\Sigma$:

$$\frac{Q_{\mathrm{theor}}}{\Sigma} = \frac{g \ln \frac{r_2}{r_1}}{T\,\omega^2} = u_{\mathrm{g}} \tag{19}$$

For a test-tube centrifuge with an angle head, $\ln(r_2/r_1)$ is exchanged for $\ln(1 + D/r_1 \cos\beta)$, where D is the inside diameter of the tube, and β is the angle between the tube and the axis of rotation.

The spin test is also a valuable tool for measuring *phase volume ratios*. The volumetric solids content, as determined by the spin test is an important parameter in determining a crude mass balance around the centrifuge. Spin tests performed under different time and g number conditions also give an indication of solids compressibility and rheology. It is also valuable to work up a sample of the solids phase after spinning to give an indication of its dry solids content after centrifugation. A spin test makes it possible to identify the nature and amount of different types of solid impurities, which may be harmful to the performance of the chosen machine.

7.4. Important Properties of the Process Streams

Particle size and *density difference* are the most important data, and in applications for sedimenting centrifuges they vary within wide limits: the particle size between 0.2 µm and 0.5 mm and the density differences between 15 and 2000 kg/m^3. The latter figure is valid for solid particles. The density differences between liquid phases are seldom larger than 300 kg/m^3. Often particle properties are unknown and/or difficult to measure. The spin test as described above often serves as a good measure.

As many as possible of the following data should be determined before the selection of a centrifuge is made [40]:

- Chemical composition of liquid and solid phases
- Densities of liquid and solid phases
- Viscosities of liquid and solid phases
- Particle size distribution
- pH
- Temperature
- Vapor pressure

Table 4. Important applications of centrifuges in the chemical and related industries

Application	Solids-ejecting	Solid wall disk	Nozzle	Decanter
Phosphoric acid			x	x
Hydrogen peroxide		x		
Plastics monomers				
caprolactam	x			
terephthalic acid				x
hexamethylenediamine	x			
Plasticizers	x			
Oil additives	x		x	x
Plastics				
polycarbonate		x		
PVC	x			x
polysterene				x
polyolefins				x
Pesticides	x			
Lignosulfonates	x		x	
Tar	x		x	x

Table 5. Some applications of centrifuges in the pharmaceutical and biotechnological industries

Application	Solids-ejecting, radial	Solids-ejecting, axial	Solid wall disk	Nozzle	Decanter	Tubular bowl	Chamber bowl	Multichamber bowl	Zonal rotors
Antibiotics	x		x		x		x		
Enzymes	x	x			x				
Amino acids				x	x				
Vaccines	x	x				x	x		x
rDNA products									
cell harvesting	x	x		x	x		x		
cell debris	x	x			x				
inclusion bodies	x	x			x				
Blood fractionation	x		x			x	x	x	
Insulin from pancreas	x	x			x		x		

Table 6. Some applications of centrifuges in environmental protection

Application	Solids-ejecting	Nozzle	Decanter
Municipal sewage			
activated sludge thickening		x	x
activated sludge dewatering		x	x
digested sludge dewatering		x	x
Industrial slop oils	x	x	x
Flue gas desulfurization			x

– Solubility of one phase in the other
– Corrosivity
– Erosivity
– Shear sensitivity
– Angle of repose
– Toxicity
– Explosion hazard

Apart from Stokes Law the *density* is important for the stress in the bowl material. *Low pH* and *high temperature* may, together with even a moderate chloride concentration, lead to difficulties with *corrosion*. The corrosive properties of process materials can be very serious but it is in many cases possible to reduce the corrosion rate by intermittent washing with caustic solution, e.g., after every 4 – 5 h of operation to re-establish a passivating layer on the metal surfaces. *Erosion* may be a problem with many inorganic crystalline materials, including silt-containing process water. This may necessitate the use of hydrocyclones to remove these particles. The *angle of repose* (see Section 3.2) is the angle a cone of solids forms if allowed to settle,

Table 7. Some fuel and energy-generation applications of centrifuges

Application	Solids-ejecting	Nozzle	Decanter	Solid wall disk	Tubular bowl	Chamber bowl
Petroleum production						
drilling mud, solids removal			x			
dehydration of crude oil	x					
de-oiling of water	x					
recovery of slop oils	x	x	x			
from coal	x	x	x			
from tar sand		x	x			
Coal dewatering			x			
Nuclear fuel production						
uranium enrichment					x	
fuel reprocessing					x	x
Gas turbine power production						
fuel oil water washing	x					
Diesel power production						
fuel oil purification	x					
Nuclear power production						
plant wash water cleaning	x					
Biodiesel production	x			x		

e.g., between the nozzles in a nozzle bowl. If it is 0° or close to 0°, the flow properties are extremely good. The most shear sensitive cells are mammalian cells. In the Centritech machines the *g* number is limited to 100, more than one order of magnitude lower than for other centrifuges. Some protein precipitates are also sensitive to shear [41], including those formed in the maturation tanks in breweries. Many bacteria are sensitive, but yeast cells are robust.

7.5. Improvement of Separability

According to Stokes' law the settling velocity improves when the density difference and the particle density increase and when the viscosity decreases. Ref. [42] lists physical, chemical and biological techniques for conditioning of fermentation broths. The most widely used are coagulation and flocculation, but also aging, pH change, and addition of enzymes or other chemicals are employed.

7.6. Pilot Machine Testing

After maximizing the separability of the feed stream, the preliminary $Q_{\mathrm{theor}}/\Sigma$ is determined by spin tests. The machine type and size are tentatively chosen on the basis of the preliminary mass balance and on the required Q_{act}. With the separation efficiency μ, taken from Section 2.2, Q_{act}/Σ can be calculated from the equation

$$\frac{Q_{\mathrm{act}}}{\Sigma} = \mu \cdot u_{\mathrm{g}} \tag{20}$$

It is often advisable to make a test with the chosen type of machine. The purpose of such a test is to obtain a more reliable Q_{act}/Σ value, but also to assess solids flow properties, emulsification and particle break-up tendencies and clogging risks. Recommended test procedures for a number of centrifuge types are described in [43].

7.7. Scale-Up

For all types of centrifugal separators, the Q_{act}/Σ value can be used for scale-up of the sedimentation performance if some general considerations are taken into account:

- The type of centrifuge is similar.
- The geometry is similar. For decanters, e.g., feed inlet position, *L*/*D* ratio, or beach angle. For disk machines, e.g., disk spacing, cone angle, and number of feed holes.
- The bowl speed gives similar linear velocity of wings etc. when treating shear-sensitive process feeds.
- The λ value (Eq. 5) is almost the same in both machines.

Table 8. Food industry applications of centrifuges

Application	Solids-ejecting, radial	Solids-ejecting, axial	Solid wall disk	Nozzle	Decanter
Dairy					
milk	x		x		
whey	x				
cottage cheese	x			x	
lactose					x
Starter cultures	x			x	
Vegetable oil refining	x		x		
Fish meal and oil	x				x
Meat products	x				x
Starch				x	x
Wine	x	x			x
Brewery	x	x		x	x
Fruit juice	x		x		x
Alcohol fermentation					
yeast separation				x	
distillery slop	x			x	x
Bakers' yeast					
yeast separation				x	
molasses separation	x			x	x
Sugar production	x	x		x	x
Organic acids	x			x	x
Vegetable protein	x			x	x
Alginates	x				
Coffee and tea	x				x

Table 9. Some agroindustry applications of centrifuges

Application	Solids-ejecting	Solid wall disk	Nozzle	Decanter
Natural rubber latex		x		
Vegetable oil extraction				
palm oil	x	x	x	x
olive oil	x	x		x
Wool wax recovery	x	x	x	x
Manure dewatering			x	

– The larger machine has no hydraulic limitations with respect to the calculated flows. Potential limitations are inlet and outlet pressure drops, paring-disk pumping capacities (upper and lower limits), and the upper and lower limits of solids flow.
– No mechanical limitations, e.g., power transmission, and, in decanters, gear-box torque. The equation for scaleup is

$$\left(\frac{Q_{\text{act}}}{\Sigma}\right)_1 = \left(\frac{Q_{\text{act}}}{\Sigma}\right)_2 \; (= \mu\, u_{\text{g}}) \tag{21}$$

where indices 1 and 2 denote the two machines. Direct calculations of u_g and μ are thus avoided, as are hindered settling effects.

Table 10. Important transport industry applications of centrifuges

Application	Solids-ejecting	Solid wall disk
Marine		
fuel oil purification	x	
lube oil purification	x	
hydraulic oil cleaning		x
Aviation fuel purification		x
Diesel crank-case gas cleaning		x

8. Applications

Sedimenting centrifuges are used in the most diverse fields. Tables 4–12 list the most important applications in various industries.

Table 11. Some metalworking applications of centrifuges

Application	Solids-ejecting	Solid wall disk	Decanter	Basket
Coolant purification	x	x		
Wash liquid cleaning	x	x		
Cutting oil cleaning	x	x		
Hydraulic oil purification	x	x		
Paint waste			x	x

The most important applications of the commonest centrifuge type, the radial solids-ejecting centrifuge, are production of milk and other beverages (Table 8) and purification of fuel and lube oil on board ships (Table 10). The decanter centrifuge is the most widely used piece of equipment for municipal sludge dewatering, its largest application. Another major decanter application in terms of number of units is vegetable oil production (Table 9), but other food areas are also important (Table 8).

Since sedimenting centrifuges are very suitable for rapid liquid – liquid separation it is not surprising that oil – water separation is the most successful area, either because of instability of products (food) or small space requirements (on board ships). Large decanters are very much in use in polymer production (Table 4). Disk-bowl and decanter centrifuges are used in the chemical industry for hundreds of small volume chemical products, which are too numerous to list here.

Large peripheral-nozzle machines are used in production of starch from various raw materials (Table 8).

A general area of use is lube oil and hydraulic oil purification.

In general, centrifuges that require manual cleaning and removal of solids (solid-bowl disk, tubular, chamber, and multichamber bowls) are becoming obsolete, except in those cases where the solids are very valuable, e.g., in the pharmaceuticals or fine-chemicals industries; where the solids content in the feed material is very low; or where cheap manual labor is available. In the latter case the simple imperforate basket centrifuge is used in the most diverse fields. Sophisticated equipment, such as chamber-bowl machines with intermittent solids removal, are becoming more and more interesting. More exotic bowl materials, (e.g., titanium) are spreading also to large centrifuges, increasing the range of applications. The hydrodynamics of decanter centrifuges in particular are becoming better understood, and this will in time lead to new applications. One trend is that centrifuges are becoming larger, but low-cost centrifuges made of lightweight materials are taking over some classical filter applications, e.g., purification of hydraulic oils.

Table 12. Some applications of centrifuges in the mining and metallurgical industries

Application	Solids-ejecting	Nozzle	Decanter
Kaolin			
classification		x	x
dewatering		x	x
Chalk			
classification		x	x
dewatering		x	x
Hydrometallurgy			
copper extraction	x		

9. References

General References

1. T. Svedberg, K. Pedersen: *The Ultracentrifuge*, Clarendon, London – Oxford 1940.
2. V. I. Sokolov: *Moderne Industriezentrifugen*, VEB Verlag Technik, Berlin 1971.
3. M. Osenga, *Diesel Progr. N. Am. Ed.* **70** (2004) no. 2, 22–23.
4. D. B. Purchas (ed.): *Solid/Liquid Separation Equipment Scale-Up*, Uplands Press Ltd, Croydon 1977.
5. H.-W. Hsu: *Separations by Centrifugal Phenomena*, John Wiley & Sons, New York 1981.
6. H. Axelsson: "Centrifugation" in C. L. Cooney, A. E. Humphrey (eds.): *Comprehensive Biotechnology*, **vol.2**, Chap. 21, Pergamon Press, Oxford 1985, pp. 325 – 346.
7. L. Svarovsky (ed.): *Solid – Liquid Separation*, 3rd ed., Butterworth, London 1990.

8. D. Rickwood (ed.): *Preparative Centrifugation. A Practical Approach*, Oxford University Press, Oxford 1992.
9. K. Heiskanen: *Particle Classification*, Chapman & Hall, London 1993.
10. M. Ungarisch: *Hydrodynamics of Suspensions*, Springer-Verlag, Berlin – Heidelberg 1993.
11. P. Hambleton, J. Melling, T. T. Salusbury (eds.): *Biosafety in Industrial Biotechnology*, Chapman & Hall, London 1994.
12. A. Rushton, A. S. Ward, R. G. Holdich: *Solid – Liquid Filtration and Separation Technology*, VCH Verlagsgesellschaft, Weinheim 1996.
13. A. Letki, R. T. Moll, L. Shapiro, *Kirk-Othmer*, 4th ed., **vol. 21**, pp. 828 – 875.
14. W. W.-F. Leung: *Industrial Centrifugation Technology*, McGraw-Hill, New York 1998.
15. H. Axelsson: "Cell Separation: Centrifugation" in M. C. Flickinger, S. W. Drew (eds.): *The Encyclopedia of Biopress Technology: Fermentation, Biocatalysis & Bioseparation*, John Wiley & Sons, New York (in press).
16. F. A. Records, K. Sutherland: *Decanter Centrifuge Handbook*, Elsevier Advanced Technology, Oxford 2001.

Specific References

17. K. Zeitsch: "Centrifugal Filtration" in [7], p. 476.
18. T. W. Simpson, *Adv. Filtr. Sep. Technol.* **15** (2002) 233–245.
19. D. B. Todd, G. R. Davies, *Filtr. Sep.* **10** (1973) 663 – 666.
20. J. Graham, "Separation in Zonal Rotors" in [8], pp. 315 – 350.
21. H. Roth, D. Rickwood: "Centrifuges and rotors" in [8], pp. 43 – 76.
22. D. L. Meikrantz, S. B. Meikrantz, L. L. Macaluso, *Chem. Eng. Commun.* **188** (2001) 115–127.
23. W. Duan, C. Song, Q. Wu, X. Zhou, J. Zhou, *Sep. Sci. Technol.* **40** (2005) 1871–1883.
24. D. Fishlock, *Chem. World* **1** (2004) no 6, 46–49.
25. B. Madsen, *Inst. Chem. Engrs. Symp. Ser.* No. 113 (1989) 301 – 317.
26. N. F. Madsen, *Inst. Chem. Engrs. Symp. Ser.* No. 113 (1989) 281 – 299.
27. A. Lee, *Adv. Filtr. Sep. Technol.* **10** (1996) 248 – 253.
28. S. Apelman, T. Björling, *Biotech. For. Eur.* **8** (6) (1991) 356 – 358.
29. D. Rickwood: "Specifications of Ultracentrifuge Rotors", App 2 in [8], pp. 355 – 365.
30. T. Lagerstedt, O. Nåbo in C. Veret (ed.): *Flow Visualization 4*, Hemisphere, New York 1987, pp. 603 – 608.
31. O. Nåbo, T. Lagerstedt in R. Reznicek (ed.): *Flow Visualization 5*, Hemisphere, New York 1990, pp. 676 – 681.
32. O. Nåbo, C.-G. Carlsson, *Proceedings Filtech Conference*, **vol. 1**. Utrecht 1987, 1 – 7.
33. L. Bergström, C.-G. Carlsson, C. Inge, T. Lagerstedt, H. Moberg, *Appl. Sci. Res.* **53** (1994) 35 – 50.
34. I. Weibull, *Mater. Des.* **8** (1987) no. 1, 35 – 40.
35. I. Weibull, *Mater. Des.* **8** (1987) no. 2, 82 – 88.
36. C. M. Ambler, *Chem. Eng. Progr.* **48** (1952) 150 – 158.
37. C. M. Ambler, *J. Biochem. Microbiol. Technol. Eng.* **1** (1959) 185 – 205.
38. S. Aiba, S. Kitai, H. Heima, *J. Gen. Appl. Microbiol.* (1964) 243 – 256.
39. F. E. Sullivan, R. A. Erikson, *Ind. Eng. Chem.* **53** (1961) 434 – 438.
40. R. J. Wakeman, *Filtr. Sep.* **32** (1995) no. 3, 337 – 341.
41. A. I. Clarkson, M. Bulmer, N. J. Titchener-Hooker, *Bioproc. Eng.* **14** (1996) 81 – 89.
42. C. Bowden, *The Chem. Engr. (Rugby)* (1985) 415, 50 – 54.
43. F. A. Records: "Sedimenting Centrifuges," in [4], pp. 199 – 240.

Chromatographic Reactors

Jörg Fricke, Department of Chemical Engineering, University of Dortmund, Dortmund, Federal Republic of Germany

Henner Schmidt-Traub, Department of Chemical Engineering, University of Dortmund, Dortmund, Federal Republic of Germany

Motoaki Kawase, Department of Chemical Engineering, Kyoto University, Kyoto, Japan

1. Introduction 1
2. Types of Chromatographic Reactors . 2
3. Modeling of Chromatographic Reactors 7
3.1. Adsorption Isotherm and Reaction Kinetics 8
3.2. Modeling the BCR 9
3.3. Modeling Simulated Moving Bed Chromatographic Reactors 10
4. Conditions and Requirements for Preparative Chromatographic Reactors 11
5. Design of Preparative Chromatographic Reactors 13
6. Preparative Chromatographic Reactors 16
6.1. Applications of Gas Chromatographic Reactors 16
6.2. Application of Liquid Chromatographic Reactors 18
7. Analytical Chromatographic Reactors 20
7.1. Determination of Rate Constants . . . 20
7.2. Enzymatic Reactions 22
7.3. Characterization of the Stationary Phase 22
7.4. Examination of Fast Equilibrium Reactions 23
7.5. Examination of Solvent Effects 23
8. References 23

1. Introduction

Compared to the common process design of reaction and separation in series, the integration of both process steps into one unit operation may improve the reaction yield and the separation efficiency. Besides reactive distillation (→ Reactive Distillation), reactive extraction, and reactive absorption, coupling of chemical and biochemical reactions with chromatographic separation is an especially attractive integrated process for producing high-purity products. The chromatographic reactor can be defined as "a chromatographic column in which a solute or several solutes are intentionally converted, partially or totally, to products during their residence in the column". As the apparatus is employed as reactor and separator, both conversion of the starting material and separation of the components take place in the column [1]. A more general definition is to define a chromatographic reactor as unit operation in which a chemical or biochemical reaction is combined with a chromatographic separation.

In chromatographic systems, a fluid mobile phase is in constant contact with the stationary phase during its passage along the column. The mobile phase may be a gas or a liquid, and therefore a distinction between gas chromatographic reactors (GCR) and liquid chromatographic reactors (LCR) can be made. The most common stationary phase in chromatographic reactors is a solid adsorbent, in the form of particles with large internal and external specific surface areas. The stationary phase may also be a liquid coating on a solid support, or chemical substances of varying chain length and molecular mass bonded to the surface of a support. The reaction can take place in the stationary phase and/or in the mobile phase. Heterogeneous reactions can be catalyzed by the adsorbent itself or by an additional catalyst that is mixed with the adsorbent.

Chromatographic reactors are used for analytical and preparative applications. On the analytical scale, they are used for determining reaction kinetics, characterizing the stationary phase, or examing the interaction between mobile and stationary phases. On the preparative

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.c07_c01

scale, chromatographic reactors are employed for the production of chemicals and the preparation of unstable reagents for laboratory use. The use of chromatographic reactors on a preparative scale is especially attractive where a chromatographic separation has already been used for purification of the products. Laborious screening of a suitable adsorbent can then be omitted, and only the catalyst needs to be chosen. Hence, the cost of process development is significantly reduced.

The integration of reaction and chromatographic separation offers various advantages. For reversible reactions, conversion can be increased beyond chemical equilibrium by continuously removing the products from the reaction zone. Combination of two unit operations in one apparatus leads to smaller capital expenditure. In analytical applications, impurities in the reactants can be easily detected, and therefore the expense of prior preparation of the components can be reduced. The analysis of the measurements is more comfortable because no additional instrument is needed for determining the amount of the reactants and the products.

Although chromatographic reactors offer various advantages, no commercial application on a preparative scale is known so far. Nevertheless, especially in the field of enzymatic reactions, the possibility of applying a chromatographic reactor is increasingly being examined. Therefore, this new process is expected to find its first industrial applications in biochemical reactions.

2. Types of Chromatographic Reactors

Most chromatographic reactors are identical to the analogous apparatus used exclusively for separation. The only reactors that have especially been developed for the combination of reaction and separation are the Hashimoto reactor and the reversed flow reactor.

Batch Chromatographic Reactor (BCR). The simplest way to realize a chromatographic reactor is to run it in a batch mode. Figure 1 illustrates the operating principle of a batch chromatographic reactor for a reversible decomposition reaction. The starting material A is injected as a sharp pulse into the column. During its propagation along the column it is converted to the products B and C. The different adsorption of components B and C leads to different propagation velocities, and the products are separated from each other. Therefore, it is possible to totally convert the starting material into the products although the reaction is limited by chemical equilibrium. Additionally, no separation unit following the chromatographic reactor is necessary to obtain the products in high purity. The operating conditions should be chosen so to achieve complete conversion. Otherwise, an additional separation unit is necessary to recycle the starting material.

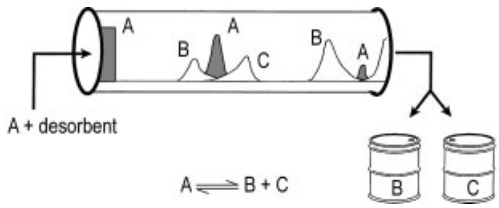

Figure 1. Schematic representation of the chromatographic batch reactor

Two fields of application have been described in the literature so far. First, the BCR can be used to analytically determine the reaction rate constants or the mechanism of a reaction and to characterize the stationary phase (see Chap. 7). The BCR can also be used for the production of chemicals (see Chap. 6).

After the invention of the chromatographic reactor in 1961 [2], [3], most experimental and theoretical studies dealt with the concept described above. Research on preparative-scale chromatographic reactors has mainly focused on continuous chromatographic reactors.

Continuous Rotating Annular Chromatograph (CRAC). By using a cylindrical annulus, the chromatographic reactor can be operated continuously. The annulus space is formed by employing two concentric cylinders. The annulus, packed with adsorbent and/or catalyst, is rotated continuously. The components are fed into the apparatus through a fixed port at the top of the annulus. Additionally, eluent is uniformly introduced over the whole cross section at the top of the apparatus. A component in the mobile phase moves in an axial direction (because

of gravity and pressure difference) and in a tangential direction (because of the rotation of the apparatus). An adsorbed molecule only moves tangentially. Because of the different adsorption of the components, their residence time in the stationary phase differs. Therefore, the components can be collected at fixed outlet ports at the bottom of the annulus (Fig. 2).

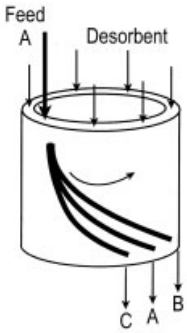

Figure 2. Schematic representation of the rotating cylindrical annulus

This principle was first introduced by MARTIN [4]. The possibility of applying this concept to chromatographic reactors was demonstrated by CARR et al. for the hydrolysis of methyl formate [5]. Other reactions that have been performed with the rotating cylindrical annulus are the enzyme-catalyzed inversion of sucrose [6] and the saccharification of starch to maltose [7].

The production rate of the rotating cylindrical annulus is comparable to that of the BCR [8]. The rotating annulus can be considered as a number of batch columns arranged in a circle and periodically charged with feed. This also explains why the performance of the two variants is the same.

A major drawback of the BCR and the CRAC is the large amount of eluent needed to operate the process. This also leads to a high dilution of the products. Another disadvantage is the poor use of the adsorbent. These drawbacks led to the development of continuous countercurrent processes.

True Moving Bed Chromatographic Reactor (TMBCR). In the true moving bed chromatographic reactor (TMBCR), also called countercurrent moving bed chromatographic reactor, the solid phase is moved in the opposite direction to the liquid flow. The main advantage of such an arrangement is the countercurrent flow, as in heat exchangers or distillation columns, that maximizes the average driving force. Thus, the efficiency with which the stationary phase is used increases. The countercurrent moving bed reactor was first mentioned by TAKEUCHI et al. [9]. On the basis of theoretical studies on a reaction of type $A \rightleftharpoons B$ an improvement in the conversion compared to equilibrium was found. For consecutive reactions of type $A \rightarrow B \rightarrow C$, the selectivity for the intermediate product in consecutive reactions can also be improved. An experimental study of the hydrogenation of 1,3,5-trimethylbenzene to 1,3,5-trimethylcyclohexane in excess H_2 confirmed the results found for reversible reactions of type $A \rightleftharpoons B$ [10].

The movement of the solid particles is difficult to realize. One reason is the unavoidable backmixing of the solid that reduces the efficiency of the process. Another problem is the abrasion of the particles caused by the movement.

Simulated Moving Bed Chromatographic Reactor (SMBCR). The invention of the simulated moving bed (SMB) concept [11] overcame the difficulties related to the movement of the solid. The valves of interconnected columns are sequentially switched in the direction of the fluid flow. This means the inlets (desorbent, feed) and outlets (extract, raffinate) are regularly changed. The time between the switching of the valves is called a cycle. At the end of the cycle, each inlet and outlet is moved one step in the direction of fluid (Fig. 3 A). Therefore, the countercurrent movement of the phases is simulated by sequentially changing the in- and outlets. When the number of columns tends to infinity and the switching time and column length tend towards zero, the performance of an SMB unit becomes identical to that of the corresponding TMB process. The principle of an SMB chromatographic reactor can be explained by considering a reversible reaction of type $A \rightleftharpoons B + C$ in an SMB apparatus consisting of four sections. The in- and outlets can be used to define four sections. Each section accomplishes a special task that is realized by the right choice of flow rates (Fig. 3 B):

1) *Section I* is located between the desorbent and the extract nodes. The flow rate is higher

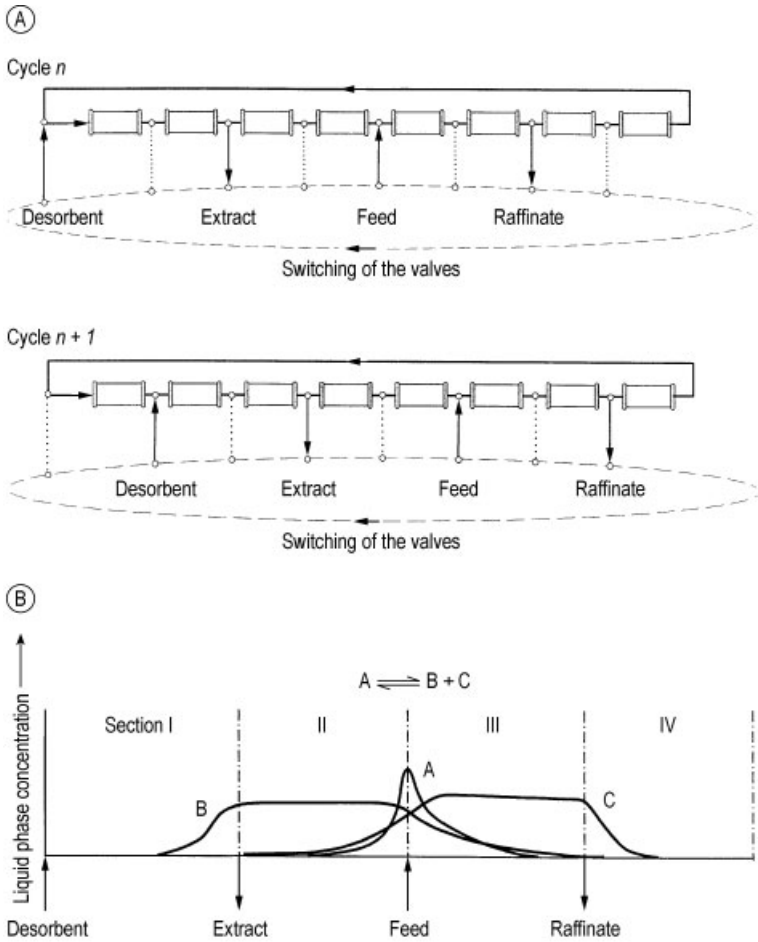

Figure 3. Schematic representation of the SMBCR
A) Configuration of the apparatus in two consecutive cycles; B) Axial profiles of the components at the end of a cycle (steady state)

than in all the other sections. This high flow rate is necessary to remove the more strongly adsorbed product (component B) from the adsorbent.

2) In *section II*, between the extract and the feed nodes, components B and C are formed. The less strongly adsorbed product (component C) is desorbed, whereas B is still held on the adsorbent. The extract stream therefore contains the more strongly adsorbed product.
3) *Section III* is also needed for the decomposition of component A. Here component B is adsorbed. Thus the fluid collected at the raffinate port has a high concentration of component C.
4) Before the fluid is recycled, it is cleaned in *section IV*. Component C is adsorbed and recycled to section III with the adsorbent.

Besides the concept described above, the SMBCR can be varied in different ways. For example, it can be run as open SMBCR, i.e., without a recycle stream [12], [13]. In this case, sometimes section IV is omitted, and the raffinate is collected at the outlet of section III.

Another variant has been developed for gas – solid reactions. Again, the solid flow is simulated by switching the valves. For the hydrogenation of 1,3,5-trimethylbenzene, a reactor consisting of three sections is employed. A mixture of adsorbent and catalyst is packed in the

column. The feed is injected in the feed section and reacts to form the product 1,3,5-trimethylcyclohexane, which is less strongly adsorbed than the starting material. The product stream is totally removed at the outlet of the feed section. The feed has to be switched to the next column in line before the starting material 1,3,5-trimethylbenzene reaches the outlet of the feed section. In the next cycle the former feed column, which now belongs to the carrier section, contains the starting material and the product. In the carrier section both components are desorbed and fed into the feed section. But it is not possible to totally remove the less strongly adsorbed starting material in this section. This is why a third section, known as the extra carrier section, is needed to obtain a high product purity. In this section the more strongly adsorbed starting material, that remained in the column, is removed by an extra carrier stream. At the end of the cycle this column should no longer contain either component. Therefore, after switching the valves, e.g., this section becomes the feed section, 1,3,5-trimethylcyclohexane can be collected at high purity (Fig. 4 A).

In the apparatus developed for the oxidative coupling of methane two modifications to that shown in Figure 4 A must be made. First, in each section two separation columns are in series to a reaction column. Splitting the reaction and separation is necessary because of the high temperature required for reaction, whereas adsorption has to be operated at lower temperatures [14]. The second modification is necessary because of the reversed reactant and product elution order. Accordingly, the starting material reaches the column outlet first. The feed is injected in the feed section and reacts to form the products ethylene and water. The feed has to be switched to the next column in line before the starting material reaches the outlet of the feed section. In the next cycle the former feed column, which now belongs to the carrier section, contains the starting material and the products. Because of the different propagation velocities, the starting material and the products are separated in the carrier section. At the end of the cycle the first separation column only contains ethylene and water. After the switching of the valves the products can be removed by an extra carrier stream (Fig. 4 B). Further details concerning the oxidative coupling of methane and the hydrogenation of 1,3,5-trimethylbenzene can also be found in Section 6.1.

Continuous Reactor for Producing High-Fructose Syrup (CRHFS). This reactor concept was developed to achieve high conversion in the case of reactions of type $A \rightleftharpoons B$, e.g., the isomerization of glucose (see Section 6.2). This can be achieved by using different columns containing the adsorbent and the catalyst. The reactor is also operated according to the concept of simulated moving bed chromatography but consists only of three sections (Fig. 5). In section III, reaction and adsorption columns alternate, whereas sections I and II consist only of adsorption columns. The switching of the in- and outlets of the adsorption columns is similar to the SMBCR, but the position of the reaction columns does not change. These columns are connected to the adsorption columns in section III and do not move into another section. In section II the starting material is separated from the product and carried back to the reactors in section III [15].

Reversed-Flow Reactor (RFR). The concept of the chromatographic reactor with reversed flow is similar to the catalytic heat regenerator reactor. In the last named, the heat capacity of the catalyst is used to act as a thermal buffer. Therefore, the feed gas can be heated. In the chromatographic reactor not heat but mass, i.e., one reactant, is stored by adsorption (Fig. 6). The principle of the reactor can be explained by considering a reaction of type $A + B \rightarrow C$. Component B is adsorbed whereas components A and C are not adsorbed on the catalyst. The adsorbable reactant is continuously fed into the reactor via a sidestream and is adsorbed on the stationary phase. Component A is introduced at one end of the reactor and reacts with the adsorbed reactant to form product C. Before component B reaches the outlet, the flow is reversed, so that reactant A is introduced on the other side of the reactor and the concentration wave of component B propagates in the other direction.

Advantages of this reactor are the total conversion of the reactants, its insensitivity to fluctuation of the input conditions (flow rate, concentration), and simpler control of the reactor.

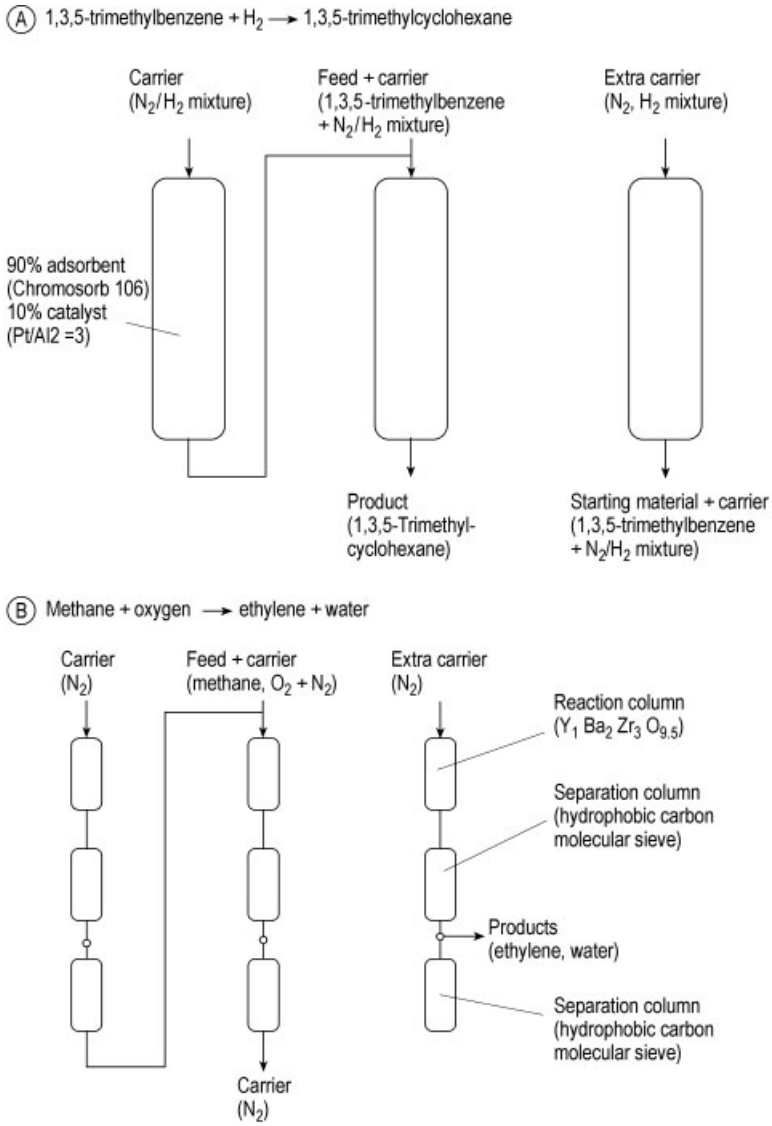

Figure 4. Schematic representation of the multi column SMBCR used by CARR et al.
A) Hydrogenation of 1,3,5-trimethylbenzene; B) Oxidative coupling of methane

A practical application of this concept has been found in the selective catalytic reduction of nitrogen oxides in stack gas with ammonia:

$$4\,NO + 4\,NH_3 + O_2 \longrightarrow 4\,N_2 + 6\,H_2O$$

Ammonia is selectively adsorbed on a stationary phase which also acts as a catalyst. By periodically reversing the flow, high conversion of NO (>99 %) and small NH_3 slip values (<1 mL/m^3) can be obtained. A list of conditions that should be fulfilled for the operation of a reversed flow reactor for the selective reduction of nitrogen oxides using ammonia has been given by AGAR and RUPPEL [16]:

- The adsorption must be selective and strong enough to yield reasonable cycle times.
- The rate of adsorption must be comparable to the rate of the reaction.
- The rate of the desorption must be slow.
- The reaction must be irreversible and free of side reactions.

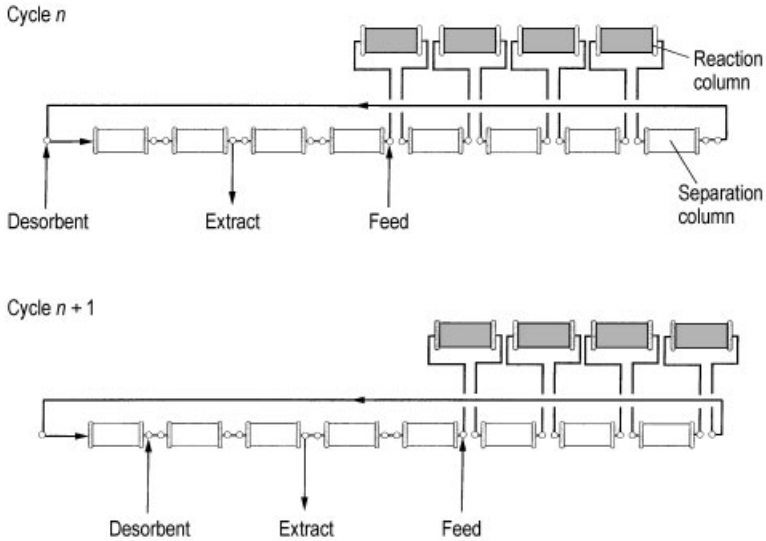

Figure 5. Schematic representation of the CRHFS; configuration of the apparatus at two consecutive cycles

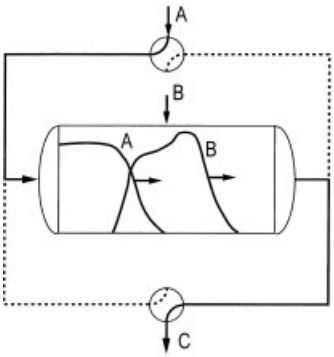

Figure 6. Schematic representation of the reversed flow reactor

3. Modeling of Chromatographic Reactors

Different models to describe chromatographic reactors have been published. They can be classified according to the modeling approach and the effects that are taken into account (Fig. 7). A chromatographic column can be described in two different ways. In the first approach the column is divided into theoretical plates that are connected in series. The second approach is based on the differential mass balance. An SMB process can also be described in two different ways. It can be approximated to be an equivalent true moving bed reactor or described as different fixed columns connected to each other, whereby the in- and outlets are periodically shifted.

Models of different complexity can be built. A distinction is made between models that take mass transfer resistance into account and models in which equilibrium between the solid and the fluid phase is assumed. By employing the model of axial dispersion the axial mixing along the column can be described. In chromatography two main mechanisms contribute to axial dispersion: molecular diffusion and eddy diffusion. These can be lumped together using a single axial dispersion coefficient. Other parameters that can be used to develop a more detailed model are intraparticle diffusion and the adsorption – desorption kinetics. The adsorption on the stationary phase can be described by an adsorption isotherm, which can have a linear or nonlinear relationship. Depending on the catalyst used, the reaction can either take place in the fluid phase or in the stationary phase.

3.1. Adsorption Isotherm and Reaction Kinetics

Various approaches for modeling the relationship between the equilibrium concentration of

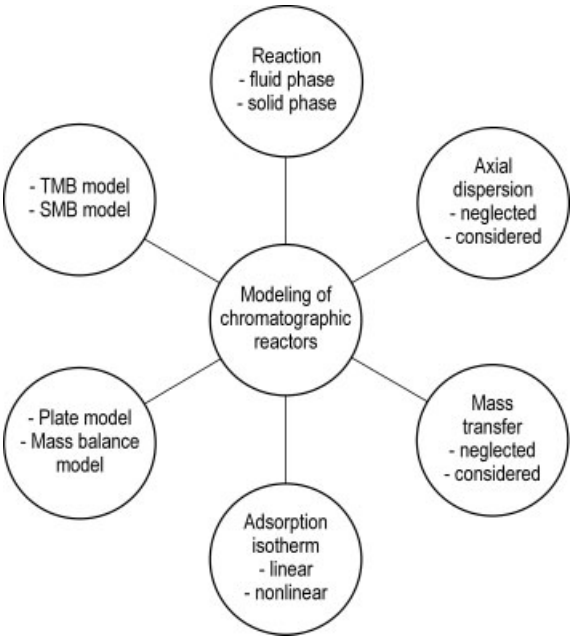

Figure 7. Modeling of chromatographic reactors

a component in the stationary and the mobile phases can be found in the literature [17].

For small concentrations, this relationship can be sufficiently described by a linear isotherm:

$$q_i = K_{\mathrm{H},i}\, c_{\mathrm{p},i} \tag{1}$$

where q_i is the concentration of component i in the stationary phase, $K_{\mathrm{H},i}$ is the Henry coefficient and $c_{\mathrm{p},i}$ is the mobile-phase concentration in equilibrium with the solid-phase concentration.

Nonlinear adsorption can be described by the competitive Langmuir isotherm model:

$$q_i = \frac{K_{\mathrm{H},i}\, c_{\mathrm{p},i}}{1 + \sum\limits_{j=1}^{n} b_j\, c_{\mathrm{p},j}} \tag{2}$$

where n is the number of components in the system and b_j is the ratio of the rate constants of adsorption and desorption.

For modeling the reaction kinetics, the reaction rate is given as a function of the concentrations of the starting material and, for a reversible reaction, of the products. For the formulation of the rate equations for chemical reactions, see → Principles of Chemical Reaction Engineering and [18].

3.2. Modeling the BCR

The Model of an Ideal Chromatographic Reactor. The model of an ideal chromatographic reactor can be used to examine the influence of the reaction kinetics on the reactor behavior. The following assumptions must be fulfilled [1]:

- The products are instantaneously separated from the reactant(s).
- Mass transfer resistance can be neglected. Additionally, the adsorption – desorption process is fast compared to chemical reaction rate. This means the overall reaction rates are controlled by chemical kinetics.
- Peak spreading and axial dispersion can be neglected.
- The adsorption isotherms are linear.
- The column is homogeneously packed and is operated isothermally.
- The fluid velocity is constant.

The resulting mass balance equation is as follows:

$$\nu_{\mathrm{fluid}} \frac{\partial c_i}{\partial x} + \frac{\partial c_i}{\partial t}\left(1 + K_{\mathrm{H},i}\right) = \nu_i r \tag{3}$$

where ν_{fluid} is the fluid phase flow velocity, c_i is the concentration of component i in the fluid phase, ν_{i} is the stoichiometric coefficient, and r is the reaction rate. The equation can be solved by using the apparent propagation velocity and the method of characteristics [19]. The model of the ideal chromatographic reactor can be used to identify fields of application where a chromatographic reactor gives considerable advantages over conventional reactors. The determination of the reaction kinetics described in Section 7.1 is also often based on the model of the ideal chromatographic reactor. In practice it is difficult to achieve an ideal chromatographic reactor. Nevertheless, its behavior can be approached if the following conditions are met [20]:

- The reaction kinetics are slow compared to adsorption and diffusion processes.
- Particles of small diameter are used for the column packing.
- A column with high efficiency is used.
- The starting material is injected as a short, sharp pulse at low concentration to fulfill the assumption of a linear adsorption isotherm.
- The temperature is well controlled over the whole column.
- The column is carefully packed, and pumps with small flow rate fluctuation are used.

Plate models are based on the assumption that a chromatographic column can be divided into a series of interconnected theoretical plates. In plate models, the axial dispersion is accounted for by the number of plates. The higher the plate number of a column, the smaller is the axial dispersion and the higher is the column efficiency. Two different types of plate models can be used, the Craig model and the Martin and Synge model ($\rightarrow$ Basic Principles of Chromatography). For the modeling of chromatographic reactors only the Martin and Synge model, which neglets mass transfer resistance, has been employed [21], [22].

$$\frac{\partial c_i^j}{\partial x}\left(\varepsilon V_{\text{c}} + (1-\varepsilon)\, V_{\text{c}} \frac{\partial q_i}{\partial c_i}\,|_{c^j}\right) = \dot{V}\left(c_i^{j-1} - c_i^j\right) + (1-\varepsilon)\,\nu_i V_{\text{c}} r \text{ for } j = 1 \ldots N_{\text{c}} \tag{4}$$

where ε is the bed porosity, V_{c} is the plate volume, c_i^{j} is the fluid phase concentration of component i on plate j and N_{c} is the number of theoretical plates.

In comparison to mass balance models only ordinary differential equations have to be solved. Although a plate model has advantages over these models with regard to calculation time, it has some drawbacks. The number of theoretical plates also depends on the components. Therefore, in nonlinear chromatography, using only one plate number for both components significantly reduces the accuracy of the model.

Mass-balance models are based on the differential mass-balance equation. Four models of this type can be found in the literature: the general rate model, the lumped kinetic model, the equilibrium-dispersive model, and a model taking into account mass transfer resistance but neglecting axial dispersion. In the general rate model, all effects of the adsorption and reaction process are analyzed in detail. The following effects are taken into account:

- Axial dispersion
- External mass transfer
- Intraparticle or stationary phase diffusion
- Adsorption – desorption processes

Nevertheless, in most cases a simpler model is sufficient for describing the BCR. The most common approach is the lumped kinetic model. Axial dispersion and mass transfer between fluid and solid phase are taken into account. Different expressions describing the reaction kinetics and the adsorption isotherm respectively can be implemented:

Fluid phase:

$$\begin{aligned}\frac{\partial c_i}{\partial t} &= -\nu_{\text{fluid}} \frac{\partial c_i}{\partial x} + D_{\text{ax}} \frac{\partial^2 c_i}{\partial x^2} \\ &- (1 - X_{\text{cat}})\, k_{\text{eff,ads},i} \frac{6}{d_{\text{p,ads}}} \frac{1-\varepsilon}{\varepsilon} \left(c_i - c_{\text{p,ads},i}\right) \\ &- X_{\text{cat}}\, k_{\text{eff,cat},i} \frac{6}{d_{\text{p,cat}}} \frac{1-\varepsilon}{\varepsilon} \left(c_i - c_{\text{p,cat},i}\right) \\ &+ \nu_i r_{\text{fluid}}\end{aligned} \tag{5}$$

where D_{ax} is the axial dispersion coefficient, $k_{\text{eff},i}$ are the overall mass transfer coefficients of component i, d_{p} is the particle diameter, and X_{cat} is the fraction of catalyst ($0 < X_{\text{cat}} < 1$).

Adsorbent:

$$\frac{\partial q_{\text{ads},i}}{\partial t} = k_{\text{eff,ads},i} \frac{6}{d_{p,\text{ads}}} \left(c_i - c_{\text{p,ads},i}\right) \tag{6}$$

Catalyst:

$$\frac{\partial q_{\mathrm{cat},i}}{\partial t} = k_{\mathrm{eff,cat},i}\frac{6}{d_{p,\mathrm{cat}}}\left(c_i - c_{\mathrm{p,cat},i}\right) + \nu_i r_{\mathrm{cat}} \tag{7}$$

If the catalyst is also the adsorbent ($X_{\mathrm{cat}} = 1$), Equation (6) can be dispensed with and only Equation (5) and (7) are used.

In the equilibrium dispersive model the stationary and the mobile phase are assumed to be in equilibrium, e.g., $c_{\mathrm{i}} = c_{\mathrm{p},i}$. Equation (5) is replaced by:

$$\frac{\partial c_i}{\partial t} + (1 - X_{\mathrm{cat}})\frac{1-\varepsilon}{\varepsilon}\frac{\partial q_{\mathrm{ads},i}}{\partial t} + X_{\mathrm{cat}}\frac{1-\varepsilon}{\varepsilon}\frac{\partial q_{\mathrm{cat},i}}{\partial t} = -\nu_{\mathrm{fluid}}\frac{\partial c_i}{\partial x} + D_{\mathrm{ax}}\frac{\partial^2 c_i}{\partial x^2} + \nu_i r_{\mathrm{fluid}} \tag{8}$$

The stationary phase concentration of the adsorbent $q_{\mathrm{ads},i}$ is given by the adsorption isotherm. The stationary phase concentration of the catalyst $q_{\mathrm{cat},i}$ is calculated from Equation (7), neglecting the term describing the mass transfer kinetics.

In the fourth model the dispersive term in Equation (5) is dropped. The mass transfer coefficient k_{eff} here includes the effect of the dispersion as far as k_{eff} is determined experimentally. Equations (6) and (7) are still valid.

For solving the mass balance equations initial and boundary conditions have to be specified. A common boundary condition is the Dankwerts condition [23].

3.3. Modeling Simulated Moving Bed Chromatographic Reactors

Modeling the continuous SMBCR process requires equations describing the relationship between internal and external flow rates (node balance):

Desorbent node (eluent):

$$Q_{\mathrm{D}} = Q_{\mathrm{I}} - Q_{\mathrm{IV}}$$
$$c_{i,\mathrm{D}}Q_{\mathrm{D}} = c_{i,\mathrm{I}}^{\mathrm{in}}Q_{\mathrm{I}} - c_{i,\mathrm{IV}}^{\mathrm{out}}Q_{\mathrm{IV}} \tag{9}$$

Extract draw-off node:

$$Q_{\mathrm{E}} = Q_{\mathrm{I}} - Q_{\mathrm{II}}$$
$$c_{i,\mathrm{E}} = c_{i,\mathrm{I}}^{\mathrm{out}} = c_{i,\mathrm{II}}^{\mathrm{in}} \tag{10}$$

Feed node:

$$Q_{\mathrm{F}} = Q_{\mathrm{III}} - Q_{\mathrm{II}}$$
$$c_{i,\mathrm{F}}Q_{\mathrm{F}} = c_{i,\mathrm{III}}^{\mathrm{in}}Q_{\mathrm{III}} - c_{i,\mathrm{II}}^{\mathrm{out}}Q_{\mathrm{II}} \tag{11}$$

Raffinate draw-off node:

$$Q_{\mathrm{R}} = Q_{\mathrm{III}} - Q_{\mathrm{IV}}$$
$$c_{i,\mathrm{R}} = c_{i,\mathrm{III}}^{\mathrm{out}} = c_{i,\mathrm{IV}}^{\mathrm{in}} \tag{12}$$

Although a true countercurrent movement of mobile and solid phase is very difficult to realize, the TMB model is often used to describe SMB processes. One reason is that the resulting differential equations are easier to handle and in some cases can even be solved analytically [24]. The hypothetical flow rate of the solid phase in a TMB process can be calculated from the switching time and the length of a column:

$$Q_{\mathrm{S}} = (1-\varepsilon)\,AL_{\mathrm{c}}/t^{*} \tag{13}$$

where Q_{s} is the solid flow rate, L_{c} is the column length and t^{*} is the switching time.

The internal flow rates are given by:

$$Q_j^{\mathrm{TMB}} = Q_j - Q_{\mathrm{s}}/F \tag{14}$$

By combining the node balance and a column model as described above, an SMB model can be built by periodically changing the boundary conditions. As the model for the batch column may be of arbitrary complexity, the resulting SMB models also differ with regard to the effects that are taken into account. An overview of the developed models is given in Table 1.

Using the simpler TMB model to describe an SMBCR has some drawbacks. The periodic characteristics of the SMB process cannot be described (Fig. 8). The resulting axial concentration profile is only an approximation of the profile in the middle of a cycle. The resulting product purities are different. Therefore, it is strongly recommended to use an SMB model.

4. Conditions and Requirements for Preparative Chromatographic Reactors

The first conditions that should be met in order to successfully operate a batch chromatographic reactor in preparative scale have been defined by Matsen et al. [32]:

Table 1. Modeling of simulated moving bed chromatographic reactors

Name	Type	Reaction	Adsorption isotherm	Ideality assumptions ∗	Comments	Ref.
SMB_1	SMB	various	various	DAX, MT		[25]
SMB_2	SMB	various	various	MT		[26]
SMB_3	SMB	various	various	DAX		[27]
SMB_4	SMB	$A \rightarrow B + C$	linear	DAX	numerical dispersion	[28]
SMB_5	SMB	$A \rightleftharpoons B$	linear	AD		[29]
SMB_6	TMB	$A \rightarrow B$	linear		analytical solution	[24]
SMB_7	TMB	$A \rightarrow B \rightarrow C$, $A \rightleftharpoons B$	linear		analytical solution	[9]
SMB_8	SMB, TMB	$A + B \rightleftharpoons C + D$	linear	MT	stage model	[30]
SMB_9	TMB	$A \rightarrow B$	various		analytical solution	[31]

∗ DAX = axial dispersion, MT = external mass transfer, AD = adsorption – desorption processes

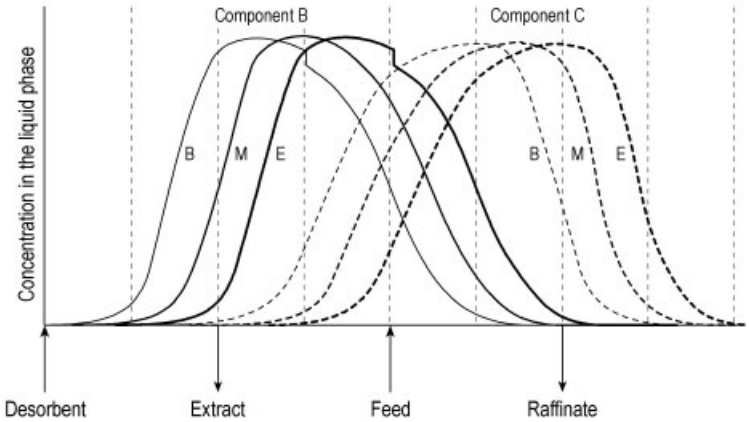

Figure 8. Axial profile of the products in an SMBCR at steady state: (SMB model): B: beginning of the cycle, M: middle of the cycle, E: end of the cycle (reaction type: $A \rightleftharpoons B + C$; type of model: SMB_1; adsorption isotherms: linear ($K_{HA} = 0.37$, $K_{HB} = 0.48$, $K_{HC} = 0.26$); reaction kinetics: $r = 1.54 \cdot 10^{-3} \cdot c_A - 1.54 \cdot 10^{-3} \cdot c_B \cdot c_C$)

1) The equilibrium constant for the reaction must be small.
2) Reaction rates should be high enough that separation of products rather than rate of reaction limits the extent of the reaction.
3) At least two products must be formed which are chromatographically separated in the reactor.
4) Reactants must not be separated in the reactor.

Examinations by SARDIN et al. showed that the above requirements can be relaxed [19]. Especially the separation of the starting materials can be overcome by adjusting the mode of operation (see also Chap. 5). As stated by the first rule given above, the application of chromatographic reactors to irreversible reactions ($A \rightarrow B$, $A \rightarrow B + C$, ...) is not as profitable as it is in the case of reversible reactions. Since no reverse reaction exists, the conversion rate cannot be increased by integrating chromatographic separation into the reactor. Therefore, in the following only reversible and consecutive reactions are considered. Besides these reactions, employing chromatographic reactors for product-inhibited reactions is expected to be advantageous compared to a conventional process of reaction and separation in series.

Reaction $A \rightleftharpoons B$. Although the chromatographic batch reactor is often used for the determination of the reaction kinetics of isomerizations, it does not offer any advantages in performing this reaction on a preparative scale. As soon as the product B is separated from the starting material A, it is consumed by the reverse reaction. Assuming first-order kinetics and linear isotherms, VILLERMAUX showed that the conversion in a batch chromatographic reactor does not

exceed that in a conventional fixed-bed reactor [33].

In contrast to operation in batch mode, conversion can be significantly increased beyond chemical equilibrium if a countercurrent reactor is employed. Based on computational studies, TAKEUCHI et al. showed that the conversion for a true moving-bed adsorber consisting of two sections is more than trebled relative to a fixed-bed reactor [9]. Experimental studies on the catalytic hydrogenation of 1,3,5-trimethylbenzene confirmed these results. The reaction was performed in a TMBCR and in a SMBCR. In both cases, experiments and simulation showed that a moving bed chromatographic reactor gave conversion of the starting material significantly exceeding the equilibrium conversion that can be attained in fixed-bed reactors [10], [29], [33].

Reaction $A \rightleftharpoons B + C$. Many studies have dealt with the reversible reaction $A \rightleftharpoons B + C$. For both the BCR and the countercurrent CR, conversion was found to exceed the limit set by chemical equilibrium. Since rules 3 and 4 as stated by MATSEN et al. are fulfilled, the use of chromatographic reactors for this type of reaction seems especially promising. Besides the general criteria given above, certain conditions concerning the adsorption isotherm should be fulfilled. It is clear that the best performance of the reactor can be achieved when the products are more or less strongly adsorbed than the starting materials. This holds for batch reactors [25], [36] and for SMB reactors [37]. Another parameter affecting the performance of SMB chromatographic reactors is the separation factor of the products. In comparison to the SMB separation process, the separation factor has to be larger in the case of the combined process. This can be explained by the formation of the products in sections II and III, i.e., close to the products nodes. Therefore, a smaller part of the apparatus can be used for separation in comparison to the separation process. In the same study, the effect of the reaction kinetics was also examined. It was shown that the use of chromatographic reactors no longer seems promising when the reverse reaction is more than one hundred times faster than the forward reaction [37].

Reaction $A + B \rightleftharpoons C + D$. Unless one of the reactants can be used as the eluent, which reduces the reaction stochiometry to $A \rightleftharpoons B + C$, separation of the reactants must be suppressed. Although separation can be compensated to a certain extent by choice of the operating conditions (see Chap. 5), it is important that the separation factor of the starting materials is small. Additionally, the products should be eluted on either side of one of the reactants [19].

Reaction $A + B \rightarrow R$, $B + R \rightarrow S$. For consecutive reactions like $A + B \rightarrow R$, $B + R \rightarrow S$ the main reason for employing reactive chromatography is to improve the selectivity. In theory, the separation of B and R should suppress the second reaction and the yield of the desired product R should be higher than in a conventional reactor.

The first study of this reaction in the BCR was performed by HATTORI et al. Based on computer simulation, they stated that a significant improvement of selectivity can be expected. The authors of this study assumed identical adsorption of the starting materials. The desired product R was more or less strongly adsorbed than components A and B. The starting materials were injected in pulses [38]. Later these results were confirmed by LIDEN and VAMLING, who also theoretically examined pulsing of components A and B, assuming that only the product R is adsorbed [39].

In contrast, a theoretical study performed by VILLERMAUX showed no improvement relative to fixed-bed reactors if component A is used as eluent while B is periodically pulsed [33], [34]. In 1993 SARDIN et al. [19] resumed the examination of this consecutive reaction in the chromatographic regime. Contrary to HATTORI et al. and LIDEN et al., they allowed different adsorptions of the starting materials. In general the concentration of component A should exceed the concentration of B. Therefore, component A must be injected as a narrow pulse, and component B as a wide pulse. Additionally, component A has to be either less or more strongly adsorbed than B and R.

An experimental investigation of the hydrolysis of glycol diacetate corroborated these theoretical results. Injecting component A at a concentration smaller than that of B makes it impossible to enhance the selectivity. Hence, no experimental proof for an advantage of using the chromatographic reactor has been

given so far. Additionally, for a realistic elution order ($K_{HA} > K_{HR} > K_{HB}$) the improvement in selectivity is rather small in the case of linear adsorption. It totally vanishes for strongly nonlinear adsorption isotherms. In conclusion, performing consecutive reactions of type $A + B \rightarrow R$, $A + R \rightarrow S$ in a chromatographic reactor does not seem to be promising [19].

5. Design of Preparative Chromatographic Reactors

The combination of reaction and separation in one unit operation makes the choice of the adsorbent, catalyst, eluent, and operating conditions more difficult. For an optimal design of a batch process, SARDIN et al. suggested a procedure consisting of six successive steps [19]. In general, the following five steps should be taken:

- Determination of the adsorption isotherm of the starting materials and products without reaction
- Determination of the reaction kinetics
- Development of a numerical model
- Validation of the model by experiments
- Choice of the operating conditions on the basis of numerical simulation

The modeling of chromatographic reactors is addressed in more detail in Chapter 3. In this chapter the choice of the adsorbent and catalyst, the design of the column, and the selection of the operating conditions are described.

Design of the Column. If a solid catalyst is used, the adsorbent and the catalyst have to be packed together. To achieve a homogeneous mixture of the particles, the particle sizes should lie in the same range. Other properties jeopardizing the homogeneity of the packing, for example, density or electrical charge, should also be taken into account. If it is not possible to avoid heterogeneous packing, the catalyst and the adsorbent may be packed in alternating sections. This arrangement should also be preferred when the lifetimes of catalyst and adsorbent differ significantly. However, the effect of the number of segments on the retention time and the conversion of the reaction has to be taken into account. For a given total reactor volume, the influence of this partitioning on the overall conversion rate is shown in Figure 9. A segmentation of 2, for example, indicates a chromatographic reactor separated into two reactor and two separation sections. A heterogeneous reactor with a high partition number operates similarly to the homogeneous reactor. A segmentation of 2 already leads to a considerable increase in the conversion (93.37 %) compared to a conventional reactor (conversion = 67.56 %). As illustrated in Figure 9, a homogeneous mixture of catalyst and adsorbent is favorable, but good results can already be obtained with small partition numbers [25].

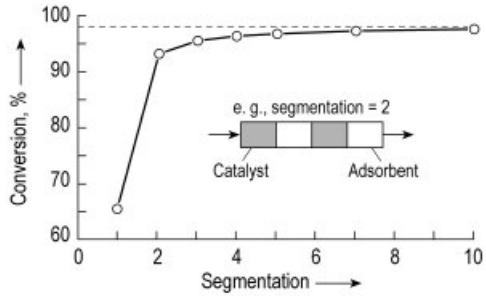

Figure 9. Comparison of homogeneous (dashed line) and heterogeneous packing (solid line) of a chromatographic column for a reaction of type $A \rightleftharpoons B + C$; (lumped kinetic model; fraction of catalyst: 0.5; adsorption isotherms: $q_{ads,A} = 1.5 \cdot c_{p,ads,A}/(1 + 0.023 \cdot c_{p,ads,A})$, $q_{ads,B} = 1.86 \cdot c_{p,ads,B}/(1 + 0.025 \cdot c_{p,ads,B})$, $q_C = 1.1 \cdot c_{p,ads,C}/(1 + 0.02 \cdot c_{p,ads,C})$; reaction kinetics: $r = 0.2\,(c_{p,cat,A} - c_{p,cat,B} \cdot c_{p,cat,C}/0.01)$

Choice of Operating Conditions.

Pressure and Temperature. Pressure and temperature have an impact on the reaction and on the separation. Increasing the temperature leads to higher reaction rates but decreases the adsorption, i.e., the retention time. In general, the known rules of equilibrium shifting are valid. For example increasing the pressure in a gas-phase reaction will shift the reaction $A \rightleftharpoons B + C$ towards A [19].

Batch Chromatographic Reactor. In batch operation the parameters injection time t_{inj}, pulse period t_{cyc}, and flow rate must be chosen. These parameters are strongly affected by the reaction kinetics, the reaction stoichiometry, and the adsorption isotherm. Based on the maximum conversion, SARDIN et al. developed simple algebraic equations to calculate the opti-

mum pulse periods [19]. These equations given in Figure 10 are based on the simplifying assumptions of linear adsorption ($K_{\mathrm{H},i}$ = Henry coefficient, see Chap. 3), plug flow, and reaction not effecting the shape of the reactant pulse.

For a reversible reaction of type $\mathrm{A}+\mathrm{B} \rightleftharpoons \mathrm{C}+\mathrm{D}$ the reactants may also be separated because of different adsorption. To increase the conversion, the time of pulse overlap should be as long as possible. This can be achieved by changing the injection mode in an appropriate way. Two possible modes of operation are conceivable. First, as suggested by Sardin et al., the concentrations of the pulses are changed. If a stoichiometric mixture is injected, this also changes the injection time. The injection time of the less strongly adsorbed reactant B can be calculated from:

$$t_{\mathrm{inj,B}} = t_0 \left(K_{\mathrm{HB}} - K_{\mathrm{HA}}\right) + t_{\mathrm{inj,A}} \tag{15}$$

The reactants can also be injected at the same concentration but at different times. One of the reactants can also be used as carrier fluid.

For the consecutive reaction $\mathrm{A}+\mathrm{B} \rightarrow \mathrm{R}$, $\mathrm{R}+\mathrm{B} \rightarrow \mathrm{S}$, the concentration of reactant A must exceed the concentration of B in the reaction zone. Assuming an equimolar feed, component A has to be injected as a sharp pulse either before or after the second reactant B (Fig. 10). The injection time $t_{\mathrm{inj,B}}$ can be estimated by using Equation (15).

Simulated Moving Bed Reactor. Successful operation of SMB chromatographic reactors depends on the correct choice of the internal flow rates and the switching time. In general, the design of these parameters can be divided into three main steps:

- General process layout: determination of the size and the number of columns
- Calculation of a first set of operating parameters by using an analytical process model
- Process optimization based on numerical simulation

In general it is preferable to run a four section SMB to enable recycling of the desorbent. In practice, it may be necessary to remove byproducts and other impurities to achieve the desired product quality. This could be done in an open-loop SMBCR process, in which the recycle stream from zone IV to zone I is eliminated. Byproducts are withdrawn with the effluent at the end of zone IV. This design still offers the advantages concerning conversion rate and product quality discussed above, but it reduces the profitability of the process because of the increased demand for fresh desorbent and the need to refine the effluent. Another argument for open-loop processes is the simplified operation of the SMB unit. If the product in the raffinate stream is not adsorbable, an open-loop configuration consisting of only three sections should be used.

The number of columns and their distribution in the different sections depends on the reaction kinetics and the efficiency of the separation. More columns are required to perform difficult separations, i.e., when the separation factor or the column efficiency is low. Especially for slow reactions it is recommended to increase the reaction zone by adding columns to section III. This can be illustrated by a case study for the SMBCR design for a sucrose inversion reaction. The initial design of the process consisting of eight columns (two columns per section) was changed in two ways:

1) The length of the column was doubled.
2) A third column of same length was added to section III of the initial design.

The resulting production rates are compared to that of the initial design in Table 2. Compared to the initial design, increasing the column length leads to a better ratio of the feed and desorbent flow rates, but the ratio of the feed flow rate to the adsorbent mass significantly decreases. Adding a third column to section III increases the values of both criteria used to calculate the production rate [40].

A first estimation of the flow rates can be made by using separation conditions developed for SMB separation processes. These can be defined on the basis of the ideal model, neglecting axial dispersion and mass transfer resistance. These assumptions allow an analytical solution of the model equations. The ratio of the net fluid flow rate and the adsorbed phase flow rate in each section is given by [41]:

$$m_j = \frac{Q_j t^* - AL_{\mathrm{C}}\varepsilon}{AL_{\mathrm{C}}\left(1-\varepsilon\right)} \tag{16}$$

By using this parameter, the following separation conditions can be defined. The inequalities given here are valid for linear adsorption:

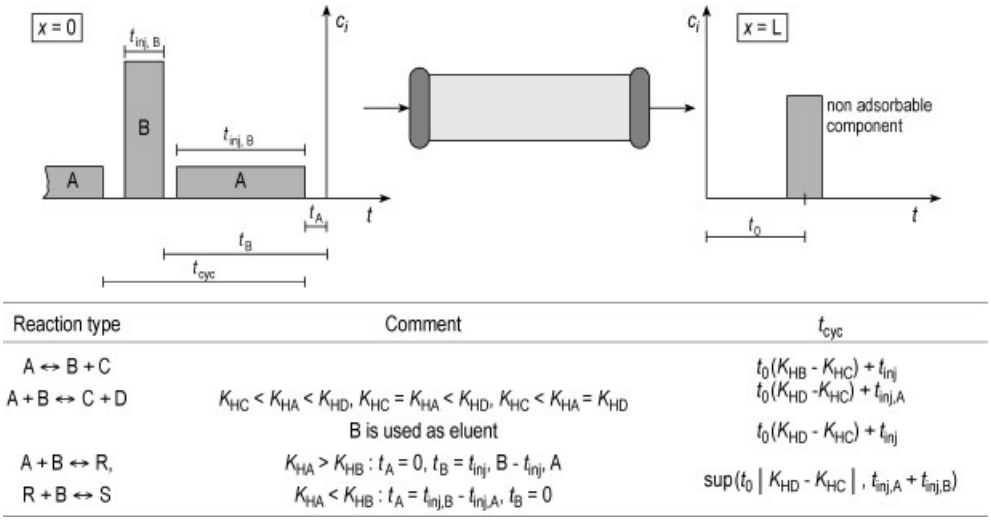

Figure 10. Optimum pulse period for BCR operation

Table 2. Comparison of the production rate for different configurations of the SMB chromatographic reactor

	Process 1	Process 2	Process 3
Process design (column length, reactor partition)	L, 2 : 2 : 2 : 2	$2L$, 2 : 2 : 2 : 2	L, 2 : 2 : 3 : 2
Feed/desorbent	100 %	150 %	144 %
Feed/adsorbent	100 %	66 %	121 %

$$H_{\mathrm{B}} \leq m_{\mathrm{I}} \tag{17}$$

$$H_{\mathrm{C}} \leq m_{\mathrm{II}} \leq m_{\mathrm{III}} \leq H_{\mathrm{B}} \tag{18}$$

$$m_{\mathrm{IV}} \leq H_{\mathrm{C}} \tag{19}$$

These inequalities give the lower limit of the flow rates in sections I and II and the upper limit of the flow rates in sections III and IV. The region of complete separation for section II and III is illustrated in Figure 11. Separation conditions for nonlinear isotherms have also been published [42].

For reasonable process optimization, numerical simulation has to be used. By modifying the switching time and the external fluid streams, the process can be optimized to maximize the conversion rate, product purity, and the feed throughput. Another goal of process optimization is to minimize the consumption of desorbent and the dilution of the products. Figure 12 illustrates the transformation of these optimization goals into a theoretical concentration profile at the end of a switching period. The conversion rate is maximized if the starting material A remains between points 2 and 3. To withdraw both products in a pure form, the concentration fronts of components B and C must not enter zone IV and zone I, respectively. The purity of the desorbent recycle stream is ensured if component C is not allowed to go beyond point 4. Similarly, the purity of the adsorbent is guaranteed if component B cannot reach the area past point 1. If these criteria are exactly met, product dilution and desorbent consumption are minimized.

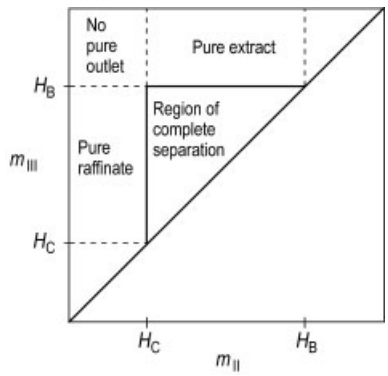

Figure 11. $m_{\mathrm{II}} - m_{\mathrm{III}}$ plane for linear isotherms

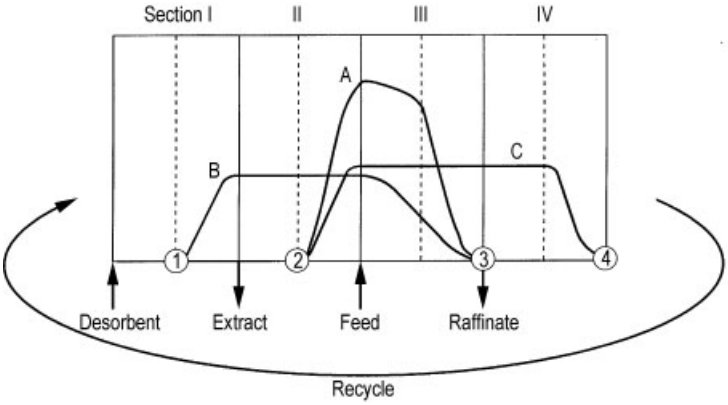

Figure 12. Optimization criteria of an SMB chromatographic reactor

Table 3. Experimental applications of gas chromatographic reactors

Reaction	Catalyst	Adsorbent	Carrier	Reactor	Ref.
Dehydrogenation of cyclohexane	platinum – alumina supported catalyst	alumina	He	BCR	[22], [32], [33], [43–45]
Dehydroisomerization of *n*-butane to isobutene	zeolites (modified with Cd, Pt)	zeolites	N_2	BCR	[46]
Ammonia synthesis	promoted iron oxide	molecular sieves, alumina	N_2	BCR	[47]
Cracking of cumene	silica – alumina	silica – alumina	H_2	BCR	[48]
Dehydrogenation of cyclohexane	platinum – alumina supported catalyst	alumina	He	CRAC	[5]
Hydrogenation of 1,3,5-trimethyl-benzene	Pt-Al_2O_3	Al_2O_3	N_2, H_2	TMBCR	[10], [49], [50]
Oxidation of carbon monoxide	activated alumina	activated alumina	N_2	TMBCR	[51]
Oxidative coupling of methane	Sm_2O_3, $Y_1Ba_2Zr_3O_{9.5}$, $Y_1Ba_2Ge_1O_{3.5}/Al_2O_3$	molecular sieve, activated charcoal, zeolite with high silica/alumina ratio	N_2	SMBCR	[52], [53]
Hydrogenation of 1,3,5-trimethyl-benzene	activated platinum on alumina	Chromosorb 106 (porous polymer adsorbent, cross-linked polystyrene)	N_2, H_2	SMBCR	[14], [35]
Reduction of nitrogen oxides	fine monoliths	fine monoliths	N_2, O_2	RFR	[55]

6. Preparative Chromatographic Reactors

6.1. Applications of Gas Chromatographic Reactors

Besides the applications described below an overview of reactions performed in gas chromatographic reactors is given in Table 3.

Oxidative Coupling of Methane. Ethylene can be produced by the oxidative coupling of methane:

$$2\,CH_4 + 0.5\,O_2 \longrightarrow C_2H_6 + H_2O$$

$$C_2H_6 + 0.5\,O_2 \longrightarrow C_2H_4 + H_2O$$

In a fixed-bed reactor the methane to oxygen ratio must be greater than 20 : 1 to suppress side reactions, primarily the complete oxidation of methane, but this also decreases the conversion of methane. To overcome the small conversion that can be reached in a single fixed-bed tubular reactor, a simulated moving bed chromatographic reactor can be used [53], [54]. Hydrophobic carbon molecular sieve and $YBa_2Zr_3O_{9.5}$ have proved to be a suitable adsorbent and catalyst, respectively [52].

In the first cycle, excess methane is fed into the reaction column. The unconverted reactants are adsorbed in the separation columns. Before methane reaches the outlet of the second separation column, the feed is switched. In the next cycle, methane, desorbed by a carrier gas, is fed into the reaction column. The switching time must be chosen so as to avoid methane break-

through but enable complete regeneration of the adsorption column in the carrier section. The ratio of methane to oxygen in the make-up feed is rather small. According to computer simulation, a methane to oxygen ratio of ca. 1.4 gives the maximum yield of ethane and ethylene [52]. Accordingly, high methane conversion can be achieved in steady state operation. The more strongly adsorbed product can be collected at the outlet of the first separation column.

Hydrogenation of 1,3,5-Trimethylbenzene. The hydrogenation of 1,3,5-trimethylbenzene has been performed in a TMBCR [10], [49], [50] and in an SMBCR [14], [35]. The reaction is equilibrium-limited and no side reactions occur. For both reactor types, the conversion and attainable purities significantly exceed the corresponding values of a fixed-bed reactor. Although the conversion in both processes is nearly the same, using an SMBCR is preferrable. Besides avoiding the problems connected with solid transport, a smaller apparatus can be used [14].

Preparation of Unstable Reagents. On laboratory scale the GCR can be used to produce unstable reagents directly before they are used for synthesis or other reactions. Reactive gas chromatography is useful for the cracking of dicylcopentadiene to cyclopentadiene. A trap at the column outlet is used to condense dicylcopentadiene.

Anhydrous formaldehyde can be produced by depolymerization of trioxane. The gas chromatographic reactor leads to nearly total conversion of the reactant. The product pulses at the column outlet can be directly used in Grignard reactions [20].

6.2. Application of Liquid Chromatographic Reactors

Various reactions, especially in the field of biochemistry, have been performed in chromatographic reactors. A summary of various applications of liquid chromatographic reactors can be found in Table 4.

Isomerization of Glucose. A practical application of running an isomerization reaction in a chromatographic reactor was demonstrated by Hashimoto et al. The reactor concept for the isomerization of glucose is described in Chapter 2. The reaction is catalyzed by immobilized glucose isomerase. The Ca^{2+} form of an ion exchanger is used as adsorbent.

The chromatographic reactor can be compared to two conventional processes. In the first process, the plug-flow reactor is combined with batch chromatographic separation. In the second, the SMB process is used for separation. The performance of the processes is evaluated in terms of the specific desorbent consumption, i.e., the ratio of desorbent flow rate to feed flow rate. Separating glucose and fructose by SMB is superior to the batch process, and producing syrup containing 45 – 65 % fructose using a chromatographic reactor requires the smallest specific desorbent consumption.

Nevertheless a problem in using this process for the production of high-fructose syrup should also be mentioned. For operating a glucose isomerase reactor, Mg^{2+} should be supplied with the solvent to activate and stabilize the enzyme. The Mg^{2+} ions may replace Ca^{2+} ions on the adsorbent, which significantly decreases the separation of fructose and glucose. Therefore, it should be checked if running this process with the Mg^{2+} form of an ion exchanger also offers advantages compared to conventional processes [15].

Inversion of Sucrose. The first known application of the chromatographic reactor in industry is the inversion of sucrose to form glucose and fructose [57]. The reaction is catalyzed by an ion-exchange resin. The products are separated by means of their different adsorption on the ion-exchange resin. Water is used as eluent. The process, developed by Boehringer Mannheim, is run in a batch mode, that is, sucrose pulses are injected at the column inlet. At the column outlet, two fractions with a high fructose and glucose concentration, respectively, can be collected. Afterwards the fractions can be distilled to give highly concentrated solutions or fed to a crystallization unit. A schematic representation of the process is given in Figure 13.

The inversion of sucrose can also be catalyzed by the enzyme invertase. This reaction has been performed in a rotating cylindrical annulus [6]

Table 4. Experimental applications of liquid chromatographic reactors

Reaction	Catalyst	Adsorbent	Reactor	Ref.
Racemic resolution of amino acid esters	immobilized α-chymotrypsin and trypsin	various	BCR	[56]
Acid-catalyzed sucrose inversion	cation-exchange resin	cation-exchange resin	BCR	[57]
Production of dextran	dextransucrase	cation-exchange resin	BCR	[28], [58], [59]
Esterification of acetic acid by ethanol	Amberlyst 15	ion-exchange resin	BCR	[19], [60]
Transesterification of menthyl acetate	acidic cation-exchange resin	acidic cation-exchange resin	BCR	[19]
Hydrolysis of glycol diacetate	OH^-	activated carbon	BCR	[19]
Saccharification of starch to maltose	maltogenase	ion-exchange resin	BCR, CRAC	[7]
Hydrolysis of methyl formate	H^+	activated carbon	CRAC	[5], [61]
Production of Bisphenol A	Amberlyst 31	Amberlyst 31	SMBCR	[13]
Esterification of acetic acid by β-phenethyl alcohol	Amberlyst 15	Amberlyst 15	SMBCR	[12]
Esterification of acetic acid by ethanol	ion-exchange resin	ion-exchange resin	SMBCR	[27]
Enzyme catalyzed sucrose inversion	invertase	cation-exchange resin	SMBCR	[36]
Production of dextran	dextransucrase	cation-exchange resin	SMBCR	[36], [63–65]
Isomerization of glucose	immobilized glucose isomerase	Ca^{2+} form of Y zeolite	CRHFS	[66], [67]

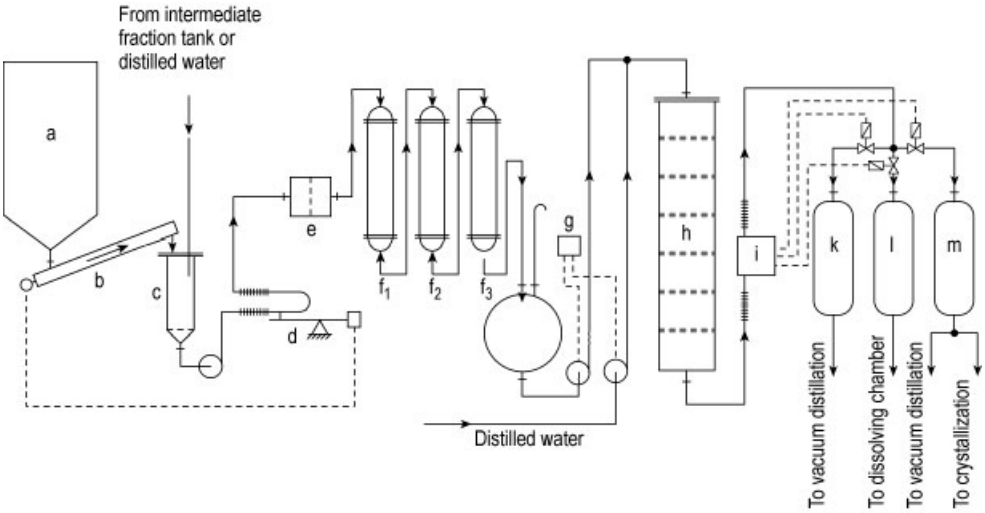

Figure 13. Schematic of a process for the inversion of sucrose
a) Sucrose silo; b) Dosing device; c) Dissolving chamber; d) Hydrostatic balance; e) Filter; f) Demineralization; g) Time switch; h) Column; i) Automatic analysis; k) Glucose tank; l) Intermediate fraction tank; m) Fructose tank [57]

and in a simulated moving bed reactor [68]. The feasibility of performing the inversion in combination with chromatographic separation has been shown in both cases.

Production of Dextran. Dextran is a polyglucose that can be synthesized from sucrose by using the enzyme dextransucrase. Fructose is formed as a byproduct:

$$n\,(C_{12}H_{22}O_{11}) \longrightarrow (C_6H_{10}O_5)_n + n\,(C_6H_{12}O_6)$$

In the pharmaceutical industry dextran is used as a blood volume expander and in the treatment of anemia. In production, a high yield of high molecular mass dextran is desired. The formation of dextran takes place by an insertion mechanism and the dextran chain is dislodged from the enzyme polymer complex by an acceptor molecule. Fructose reduces the fraction of the high molecular mass dextran by acting as an acceptor in the polymerization reaction. Therefore, the fructose concentration should be kept low. This can be done by combining the biochemical reaction and chromatographic separation. For separation an ion-exchange resin in the Ca^{2+} form can be used. The enzyme is supplied with the eluent.

The yield of dextran with molecular mass in excess of 157 000 Da can be increased by using the batch chromatographic reactor (77 %). In a conventional fermentation process only 43.7 % of high molecular mass dextran can be produced. Additionally, fructose can be obtained in high purity [28], [58], [59], [65].

The biosynthesis of dextran can also be performed in an SMBCR consisting of 12 columns. The enzyme is fed into the unit together with the adsorbent. In continuous operation, over 60 % of the produced dextran has a molecular mass exceeding 200 000 Da [63–65].

Nevertheless, some precautions must be taken when employing the SMBCR for synthesizing dextran. First of all particles with an average size not smaller than 450 µm should be used to avoid high pressure buildup in the system. Side reactions can be minimized by previous purification of sucrose. Additional calcium nitrate must be added to the system to overcome the Ca^{2+} loss of the adsorbent during operation.

Production of l-Methionine. The advantage of integration of enzymatic reaction and chromatographic separation has also been shown by using the production of L-methionine as a model reaction. L-Methionine is formed by the enantioselective deacylation of *N*-acetyl-DL-methionine. The reaction is reversible and is inhibited by the product L-methionine and the byproduct acetic acid. It is of special interest because in industry the enantioselective deacylation of *N*-acylated amino acids is a major route to enantiomerically pure L-amino acids. The enzyme, aminoacylase, can be immobilized on an ion-exchange resin that also acts as stationary phase for the separation of the products and the unconverted *N*-acetyl-D-methionine. A sodium chloride solution is used as eluent.

Experiments in the BCR showed complete conversion of *N*-acetyl-DL-methionine. After the reaction takes place in the first 5 – 10 % of the reactor, the components are chromatographically separated. At the column outlet, L-methionine, acetic acid, and the unconverted substrate D-enantiomer can be collected in high purity [69].

Esterification. Ester synthesis is an important reaction in the chemical industry. To achieve high conversion, the reversible reaction has to be followed by separation and recycling of the starting materials. By using a chromatographic reactor both processes can be combined in one unit operation. The feasibility of exceeding chemical equilibrium in a BCR was shown by Sardin et al. [19].

Among others, Kawase et al. applied the concept of SMBCR to esterifications. The production of β-phenethyl acetate was examined at different temperatures. Additionally, the reaction of methacrylic acid and *n*-butanol to form *n*-butyl methacrylate was investigated. Both reactions are catalyzed by an ion-exchange resin in the protonated form. The resin acts not only as a catalyst but also as adsorbent for the chromatographic separation. An open four section SMB configuration was used. High conversion can be reached for the production of β-phenethyl acetate at different temperatures. But for the *n*-butyl methacrylate production only a small ester yield can be obtained although the equilibrium constant is nearly identical to that of β-phenethyl acetate production. This unusual decrease in conversion can be explained by the high separation factor ($\alpha = 3.1$) of methacrylic acid and *n*-butanol. Because of the insufficient reaction rate, acid and alcohol are separated before they react to form the products [70].

Affinity Chromatographic Reactor. Many enzymes need a dissociable coenzyme for their catalytic activity. These coenzymes, for example, nicotinamide adenine dinucleotide (NAD), nicotinamide adenine dinucleotide phosphate (NADP), or adenosine 5′-phosphate (ATP), are expensive and must be regenerated or recycled.

In the affinity chromatographic reactor, the affinity between the coenzyme and the immobilized enzyme can be utilized for dynamic immobilization of the coenzyme. The resulting increase in retention time makes it possible to maximize the cycling number of the coenzyme, that is, the number of product molecules produced by one coenzyme molecule.

The affinity chromatographic reactor was successfully applied to coenzyme cycling by a single enzyme (reduction of propionaldehyde and oxidation of ethanol catalyzed by alcohol dehydrogenase [71]) and to a conjugated double-enzyme system (reduction of pyruvate catalyzed by lactate hydrogenase and oxidation of ethanol catalyzed by alcohol dehydrogenase [72]).

Aromatic Chlorination Reactions. The aromatic chlorination reactions examined by JAEGER et al. clearly demonstrate the enhancement of reactivity and selectivity in a CR. In this reaction an organic substrate, *n*-alkylphenyl ether, interacts with a water-soluble reagent, Cl^-. First the ether is immobilized on a reversed-phase chromatography column. After injection of chlorine water and gradient elution with acetonitrile water, the ratio of *para* and *ortho* products was increased relative to the homogenous reaction. Besides higher regioselectivity, also higher substrate selectivity was found for competitive monochlorination, when the reactants were of different hydrophilic/lipophilic character [73]. A similar result was also found for the base-catalyzed hydrolysis of *p*-nitrophenyl esters [74].

7. Analytical Chromatographic Reactors

7.1. Determination of Rate Constants

Inert-Standard Method. The inert-standard method is most often used for determining reaction rate constants. Assuming ideal chromatographic reactor behavior, the ratio of reactant at the column outlet $n_{\mathrm{r,out}}$ and the amount injected $n_{\mathrm{r,in}}$ is given by:

$$n_{\mathrm{r,out}}/n_{\mathrm{r,in}} = \exp\left[-\left(k_{\mathrm{m}}t_{\mathrm{m}} + k_{\mathrm{s}}t_{\mathrm{s}}\right)\right] \tag{20}$$

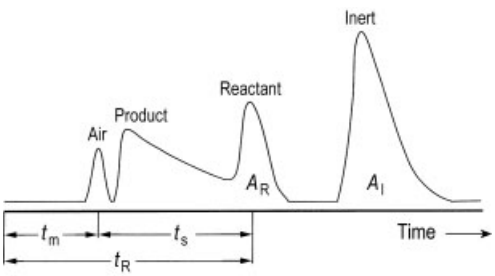

Figure 14. Illustration of a chromatogram for the reaction $A \rightleftharpoons B$ in the presence of an inert component

where t_{m} and t_{s} are the residence times of the reactant and k_{m} and k_{s} are kinetic rate constants for forward reaction in the mobile (m) and stationary (s) phase (Fig. 14).

If a known amount of inert reference material (n_i) is added to the injected feed, the peak areas A_{R}, A_{S} can be evaluated to determine an apparent kinetic rate constant k_{app}:

$$\ln\left(\frac{A_{\mathrm{R}}}{A_{\mathrm{I}}}\right) = \ln\left(\frac{n_{\mathrm{R,in}}/S_{\mathrm{R}}}{n_{\mathrm{I}}/S_{\mathrm{I}}}\right) - k_{\mathrm{app}}t_{\mathrm{R}} \tag{21}$$

where S is the detector sensitivity and t_{R} is the total column residence time of the reactant. By measuring k_{m} in a batch reactor, k_{s} can be determined from:

$$k_{\mathrm{app}} = \left(\frac{t_{\mathrm{m}}}{t_{\mathrm{R}}}\right)k_{\mathrm{m}} + \left(\frac{t_{\mathrm{s}}}{t_{\mathrm{R}}}\right)k_{\mathrm{s}} \tag{22}$$

By employing a void column between two packed columns, k_{m} and k_{s} can be measured simultaneously. Measurements with varying void volume and flow rate can be evaluated to determine the kinetic rate constants:

$$\frac{A_{\mathrm{p}} + A_{\nu}}{A_{\mathrm{I}}} = C_1 - C_2 \exp\left[\left(-k_{\mathrm{m},\nu}t_{\nu} + k_{\mathrm{m,c}}t_{\mathrm{c}} + k_{\mathrm{s}}t_{\mathrm{s}}\right)\right] \tag{23}$$

where the peak area of void product formation is given by A_{ν}, C_1 and C_2 are empirical constants, t_{ν} represents the residence time in the void column, and $k_{\mathrm{m},\nu}$ and $k_{\mathrm{m,c}}$ are the mobile phase kinetic rate constants in the void and the packed column, respectively [75], [76].

If the assumption of the ideal chromatographic reactor no longer holds, equations to evaluate the chromatogram taking into account adsorption – desorption kinetics or axial dispersion can be found in the literature [77].

Evaluation of the Overall Elution Profile. Although the inert-standard method is commonly used, sometimes the search for a suitable standard is time-consuming. Calculation of the peak area of the reactant may also be difficult. By applying the first statistical moment the overall elution profile can be evaluated for the determination of kinetic rate constants in the case of a pseudo-first-order chemical reaction [78]. Another method that can be applied here is the Fast Fourier Transform technique [79].

Stopped-Flow Method. For catalytic reactions, the stopped flow method can be used to lengthen the overall reaction time. Additionally, limitations by mass transfer resistance and slow adsorption – desorption processes may be overcome. The products that are formed during the stopped-flow period form sharp peaks superimposed on the continuous chromatogram (Fig. 15).

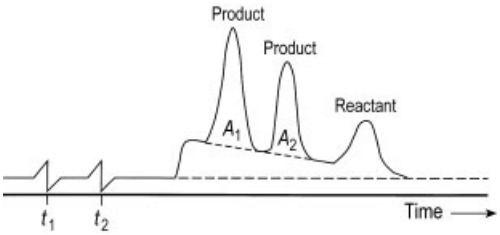

Figure 15. Illustration of a stopped-flow chromatogram for the reaction A ⇌ B

By stopping the flow twice, the resulting product peaks can be evaluated to give the reaction rate constant:

$$\ln\left(\frac{A_1}{A_2}\right) = k_s f_s\,(t_2 - t_1) + \ln\frac{\sinh\,(k_s f_s \nu_1)}{\sinh\,(k_s f_s \nu_2)} \quad (24)$$

where f_s is the fraction of the residence time of the reactant being adsorbed on the surface. The stopped-flow interval is 2ν, centered at t [1], [80].

Conditions and Requirements. The inert-standard method can be used for determining an apparent rate coefficient lying in the range between $3 \times 10^{-6}\,s^{-1}$ and $2 \times 10^{-1}\,s^{-1}$. Generally, the most accurate results are obtained when k_{app} is in the middle of this range, i.e., from $10^{-4}\,s^{-1}$ to $10^{-2}\,s^{-1}$. For slower reactions, other techniques such as the stopped-flow method must be used.

The examination of first-order or pseudo-first-order reactions is especially attractive. The experiments can be easily evaluated by using analytical solutions that determine the amount of product at the column outlet.

To distinguish between the product peak and inert peak, both components have to be sufficiently separated. This can easily be achieved with a separation factor of 1.5, but the time required for analysis is unnecessarily prolonged if the separation factor is too large. Langer and Patton suggest choosing the inert substance such that the separation factor is in the range between 1.5 and 2.0 [1].

7.2. Enzymatic Reactions

Determination of Reaction Kinetics. The chromatographic reactor has been successfully applied to determining the reaction kinetics of the hydrolysis of *o*-nitrophenyl-β-D-galactopyranoside with immobilized lactase in a microreactor. The enzyme was immobilized on porous glass beads. Besides the possibility of using already available equipment with only small modifications, the speed and ease of obtaining reproducible data are advantageous. By this method to measure enzyme activity different parameters that influence the immobilization efficiency can be easily screened [81].

Analysis of Enantioselectivity. Enantiomerically pure components can be produced by chiral resolution of the racemic mixture catalyzed by an enzyme. For process optimization the enantioselectivity must be examined under different operating conditions. This can easily be done by the method presented by Wu [82]. The number of enzyme molecules in the column has to exceed the total enantiomer input to fulfill the assumption of constant enzyme concentration.

The analysis of the peak areas at the column outlet are used to calculate the conversion of each enantiomer. The enantioselectivity is given by:

$$E = \frac{\ln\left(1 - X^{R}\right)}{\ln\left(1 - X^{S}\right)} \quad (25)$$

where X^R and X^S are the individual conversion of the R and S enantiomer, respectively.

7.3. Characterization of the Stationary Phase

Determination of Residual Oil Saturation. The amount of residual oil in a porous rock structure can be determined by injecting an ethyl acetate pulse into the well. Water acts as mobile phase, the oil in the porous rock is the stationary phase. Ethanol is formed by hydrolysis of

the tracer ethyl acetate. It is most soluble in water, whereas the tracer is soluble in both phases. After reversing the flow, the difference between ethanol and ethyl acetate reaching the well head is measured. As the difference corresponds to the residence time in the oil phase, it can be used to obtain the residual oil saturation in the rock [83], [84].

Detection of Reactive Zones in Chromatographic Columns. Reaction chromatography can be used to diagnose the existence of reactive sites in the adsorbent. An example is hydroquinone oxidation, which can be used for the detection of metal ions adsorbed on silica surfaces. The existence of active sites can be proved if decreasing the flow rate leads to increasing product peaks. The positions of the active sites can be determined by reversing the direction of flow [85]. The distribution function of active sites on a catalytic surface can be determined using stopped flow chromatography [86].

Determination of Temperature Zones. Irregularities in the temperature profile along the column can be easily detected and located in the chromatogram. The increased product formation in the hot zone caused by higher reaction rate can be identified as an extra peak in the product curve. The sensitivity of this method depends on the difference of the activation energy and the differential molar heat of vaporization of the reactant from solution. The hot zone can be most easily detected when the difference is high and the retention times of reactant and product differ significantly. A practical application of this method has been demonstrated for trioxane depolymerization [1].

Determination of Void Zones. Void zones also lead to irregularities in the chromatogram. In a void zone the reactant and the product are not separated from each other. Accordingly, the void zones can be identified as a small peak in the chromatogram. The dissociation of endodicyclopentadiene can be used to determine heterogenities in the column packing [87].

7.4. Examination of Fast Equilibrium Reactions

For fast equilibrium reactions like isomerization, enantiomerization, and intramolecular bond rotation, separation is affected by simultaneous reaction. The degree of separation depends on the reaction rate. For very fast interconversion, only one peak will appear at the column outlet. Two peaks can be observed with decreasing reaction rate. The retention time of the two peaks tends towards the corresponding elution of the pure components, the slower the isomerization takes place [88]. Examples for kinetic studies of this reaction are the separation of tautomers of β-diketoesters [89], *cis – trans* proline isomerization [90], and the separation of alpha and beta anomers of muramyldipeptide I [91].

To achieve separation, operating conditions such as temperature, mobile phase composition, and flow rates, must be chosen so as to decrease the interconversion rate. MORIYASU et al. achieved the separation of keto and enol tautomers by lowering the temperature [92].

7.5. Examination of Solvent Effects

To examine solvent effects with a gas chromatographic reactor, the solvent must be impregnated on a solid support. The reaction rate can be determined by measuring the appearance of product and the disappearance of starting material at the column outlet. Additionally solute – solvent interactions can be examined. Examples given in the literature include the determination of infinite-dilution activity coefficients.

Carrying out the chemical reaction does not lead to contamination of the solvent. This means it can be repeatedly used for other reactants or reactions. By preparing columns with different stationary phases, different solvents can be screened and the solvent choice can be optimized. Another advantage of the application of the GCR is the near complete suppression of side reactions by choosing appropriate operating conditions. Side reactions that do still occur can be easily detected. One example of the GCR being used for analyzing solvent effects is the investigation of the dissociation of endodicyclopentadiene to cyclopentadiene [93]. It was

found that the reaction rate increases with increasing liquid-phase polarity which, however, favors side reactions.

The effect of different solvents on reaction kinetics has also been examined for the thermal dissociation of several isomeric methylcyclopentadiene dimers [94].

8. References

1. S. H. Langer, J. E. Patton: "Chemical reactor applications of the gas chromatographic column," in J. H. Purnell (ed.): *New Developments in Gas Chromatography,* Wiley-Interscience, New York 1973, 293 – 373.
2. ESSO Research and Engineering Co., US 2 976 132, 1961 (J. A. Dinwiddie, W. Morgan).
3. ESSO Research and Engineering Co., CA 631 882, 1961 (E. M. Magee).
4. A. J. P. Martin, *Discuss. Farraday Soc.* **7** (1949).
5. R. W. Carr: "Continuous reaction chromatography" in G. Ganetsos, P. E. Barker (eds.): *Preparative and Production Scale Chromatography,* Marcel Dekker, New York 1993, pp. 421 – 447.
6. M. R. Sarmidi, P. E. Barker, *Chem. Eng. Sci.* **48** (1993) 2615 – 2623.
7. M. R. Sarmidi, P. E. Barker, *J. Chem. Tech. Biotechnol.* **57** (1993) 229 – 235.
8. C. Heuer, H. Kniep, T. Falk, A. Seidel-Morgenstern, *Chem. Ing. Techn.* (1997) 1535 – 1546.
9. K. Takeuchi, T. Miyauchi, Y. Uraguchi, *J. Chem. Eng. Japan* **11** (1978) 216 – 220.
10. B. Fish, R. W. Carr, *Chem. Eng. Sci.* **44** (1989) 1773 – 1783.
11. Universal Oil Products Co., US 2 985 589, 1961 (D. B. Broughton).
12. M. Kawase, T. B. Suzuki, K. Inoue, K. Yoshimoto, K. Hashimoto, *Chem. Eng. Sci.* **51** (1996) 2971 – 2976.
13. M. Kawase, Y. Inoue, T. Araki, K. Hashimoto, *Catal. Today* **48** (1999) 199 – 209.
14. M. C. Bjorklund, R. W. Carr, *Catal. Today* **25** (1995) 159 – 168.
15. K. Hashimoto, S. Adachi, H. Noujima, Y. Ueda, *Biotechnl. Bioengng.* **25** (1983) 2371 – 2393.
16. D. W. Agar, W. Ruppel, *Chem. Eng. Sci.* **43** (1988) 2073 – 2078.
17. G. Guiochon, S. D. Shirazi, A. M. Katti: *Fundamentals of preparative and nonlinear chromatography,* Academic Press, New York 1994.
18. H. Bisswanger: *Enzymkinetic,* VCH Verlagsgesellschaft, Weinheim 1994.
19. M. Sardin, D. Schweich, J. Villermaux: "Preparative fixed-bed chromatographic reactor," in G. Ganetsos, P. E. Barker (eds.): *Preparative and Production Scale Chromatography,* Marcel Dekker, New York 1993, pp. 477 – 521.
20. J. Coca, G. Adrio, C. Y. Jeng, S. H. Langer: "Gas and liquid chromatographic reactors," in G. Ganetsos, P. E. Barker (eds.): *Preparative and Production Scale Chromatography,* Marcel Dekker, New York 1993, pp. 449 – 475.
21. T. Falk, A. Seidel-Morgenstern, *Chem. Eng. Sci.* **54** (1999) 1479 – 1486.
22. D. Schweich, J. Villermaux, *Ind. Eng. Chem. Fundam.* **17** (1978) 1 – 7.
23. P. V. Danckwerts, *Chem. Eng. Sci.* **2** (1953) 1 – 13.
24. K. Takeuchi, Y. Uraguchi, *J. Chem. Eng. Japan.* **9** (1976) 246 – 248.
25. M. Meurer, U. Altenhöner, J. Strube, H. Schmidt-Traub, *J. Chromatogr.* **769** (1997) 71 – 79.
26. M. Kawase, Y. Inoue, T. Araki, K. Hashimoto, *Catal. Today* **48** (1999) 199 – 209.
27. M. Morbidelli, A. Kruglov, B. Neri, D. Gelosa, M. Mazotti, *Chem. Eng. Sci.* **51** (1996) no. 10, 1827 – 1836.
28. I. Zafar, P. E. Barker, *Chem. Eng. Sci.* **43** (1988) 2369 – 2375.
29. A. K. Ray, R. W. Carr, *Chem. Eng. Sci.* **50** (1995) 3033 – 3041.
30. V. B. Pai, J. L. Gainer, G. Carta: "Simulated moving bed chromatographic reactors" in M. D. LeVan (ed.): *Fundamentals of Adsorption,* Kluwer Academic Publishers, Dordrecht 1996, 733 – 740.
31. D. Altshuller, *Chem. Eng. Commun.* **19** (1993) 363 – 375.
32. J. M. Matsen, J. W. Harding, E. M. Magee, *J. Phys. Chem.* **69** (1965) 522 – 527.
33. J. Villermaux: "The chromatographic reactor," *NATO Adv. Study Series* (1981) 539 – 588.
34. D. Schweich, J. Villermaux, *Chem. Eng. J.* **24** (1982) 99 – 109.
35. A. K. Ray, R. W. Carr, *Chem. Eng. Sci.* **50** (1995) 2195 – 2202.

36. G. Ganetsos, P. E. Barker, J. N. Ajongwen: "Batch and continuous chromatographic systems as combined bioreactor- seperators," in G. Ganetsos, P. E. Barker (eds.): *Preparative and Production Scale Chromatography,* Marcel Dekker, New York 1993, pp. 375 – 394.
37. J. Fricke, M. Meurer, J. Dreisörner, H. Schmidt-Traub, *Chem. Eng. Sci.* **54** (1999) 1487 – 1492.
38. T. Hattori, Y. Murakami, *J. Catal.* **10** (1968) 114 – 122.
39. G. Liden, L. Vamling, *Chem. Eng. J.* **40** (1989) 31 – 37.
40. M. Meurer, U. Altenhöner, J. Strube, H. Schmidt-Traub, *Chem. Ing. Tech.* **69** (1997) 1627 – 1631.
41. M. Mazotti, G. Storti, M. Morbidelli, *J. Chromatogr.* **769** (1997) 3 – 24.
42. A. Gentilini, C. Migliorini, M. Mazotti, M. Morbidelli, *J. Chromatogr.* **805** (1998) 37 – 44.
43. D. Schweich, J. Villermaux, *Ind. Eng. Chem. Fundam.* **21** (1982) 47 – 51.
44. D. Schweich, J. Villermaux, *Ind. Eng. Chem. Fundam.* **21** (1982) 51 – 55.
45. D. Schweich, J. Villermaux, M. Sardin, *AIChE J.* **26** (1980) 477 – 486.
46. M. R. Sad, C. A. Querini, R. A. Comelli, N. S. Figoli, J. M. Parera, *App. Catal. A: General* **146** (1996) 131 – 143.
47. B. D. Unger, R. G. Rinker, *Ind. Eng. Chem. Fundam.* **15** (1976) 225 – 227.
48. Y. Murakami, T. Hattori, T. Hattori, *J. Catal.* **10** (1968) 123 – 127.
49. B. Fish, R. W. Carr, R. Aris, *Chem. Eng. Sci.* **41** (1986) 661 – 668.
50. T. Petroulas, R. Aris, R. W. Carr, *Chem. Eng. Sci.* **40** (1985) 2233 – 2240.
51. K. Takeuchi, Y. Uraguchi, *J. Chem. Eng. Japan,* **10** (1977) 455 – 460.
52. A. V. Kruglov, M. C. Bjorklund, R. W. Carr, *Chem. Eng. Sci.* **51** (1996) 2945 – 2950.
53. A. L. Y. Tonkovich, R. W. Carr, *Chem. Eng. Sci.* **49** (1994) 4657 – 4665.
54. A. L. Y. Tonkovich, R. W. Carr, *Chem. Eng. Sci.* **49** (1994) 4647 – 4656.
55. D. W. Agar, W. Ruppel, *Chem. Eng. Sci.* **43** (1988) 2073 – 2078.
56. J. Kalbe, H. Höcker, H. Berndt, *Chromatographia* **28** (1989) 193 – 196.
57. K. Lauer, *Starch/Stärke* **32** (1980) 11 – 13.
58. P. E. Barker, I. Zafar, R. M. Alsop: "A novel method for the production of dextran and fructose" in G. W. Moody, P. B. Baker (eds.): *Bioreactors and Biotransformations,* Elsevier Applied Science Publisher, Amsterdam 1987, pp. 141 – 157.
59. P. E. Barker, I. Zafar, R. M. Alsop: "Production of dextran and fructose in a chromatographic reactor- separator" in M. S. Verall, M. J. Hudson (eds.): *Separation for Biotechnology,* Ellis Horwood Ltd., Chichester 1987, pp. 127 – 151.
60. M. Sardin, J. Villermaux, *Nouveau Journal de Chimie* **3** (1979) 255 – 261.
61. B. K. Cho, R. W. Carr, R. Aris, *Chem. Eng. Sci.* **35** (1980) 74 – 81.
62. M. Kawase, T. B. Suzuki, K. Inoue, K. Yoshimoto, K. Hashimoto, *Chem. Eng. Sci.* **51** (1996) 2971 – 2976.
63. P. E. Barker, G. Ganetsos, J. Ajongwen, A. Akintoye, *Chem. Eng. J.* **50** (1992) 23 – 28.
64. G. Ganetsos, P. E. Barker, A. Akintoye, *IChemE Symp. Ser.* **118** (1990) 17 – 24.
65. P. E. Barker, G. Ganetsos: "Biochemical reaction and separation in chromatographic columns" in A. E. Rodrigues et al. (eds.): *Adsorption: Science and Technology,* Kluwer Academic Publishers, Dordrecht 1989, pp. 491 – 504.
66. K. Hashimoto, S. Adachi, Y. Shirai: "Development of new bioreactors of a simulated moving-bed type", in G. Ganetsos, P. E. Barker (eds.): *Preparative and Production Scale Chromatography,* Marcel Dekker, New York 1993, pp. 395 – 419.
67. S. Adachi, *J. Chromatogr.* **658** (1994) 271 – 282.
68. A. Akintoye, G. Ganetsos, A. E. Barker, *Trans. Inst. Chem. Eng.* **25** (1991) 2371 – 2393.
69. M. L. Jansen et al.: "Development of a Chromatographic Immobilised-Enzyme Reactor for the Production of L-Methionine" in M. L. Jansen (ed.): *Integration of Ion Exchange Chromatography with an Enzymatic Reaction,* Ph. D. Thesis, Technical University of Delyt 1996.
70. M. Kawase, Y. Masaki, J. Fricke, T. Tanigawa, K. Hashimoto: "Separation of reactants in the simulated moving-bed reactor" in *Proceedings of the Asia-Pacific Chemical Reaction Engineering Symposium 99* (1999) 263 – 268.
71. O. Miyawaki, T. Yano, *Biotechnol. Bioengn.* **39** (1992) 314 – 319.
72. O. Miyawaki: "Affinity chromatographic reactor," in G. Ganetsos, P. E. Barker (eds.): *Preparative and Production Scale Chromatography,* Marcel Dekker, New York 1993, pp. 607 – 615.

73. D. A. Jaeger, M. W. Clennan, D. E. Leyden, R. S. Shreedhara, *Tetrahedron Lett.* **28** (1987) 4805 – 4808.
74. D. A. Jaeger, M. W. Clennan, *J. Org. Chem.* **53** (1988) 3985 – 3988.
75. A. H. T. Chu, S. H. Langer, *J. Chromatogr.* **384** (1987) 231 – 248.
76. T. D. Griffith, A. H. T. Chu, S. H. Langer, *Chem. Eng. J.* **36** (1987) 73 – 91.
77. C.-Y. Jeng, S. H. Langer, *J. Chromatogr.* **589** (1992) 1 – 30.
78. C.-Y. Jeng, S. H. Langer, *Ind. Eng. Chem. Res.* **30** (1991) 1489 – 1499.
79. J. T. Hsu, U. P. Ernst, *Chem. Eng. Sci.* **45** (1990) 1017 – 1030.
80. K. F. Scott, C. S. G. Phillips, *J. Chromatogr.* **112** (1975) 61 – 70.
81. H. R. Melton, D. C. Bailey, S. H. Langer, *J. Chem. Tech. Biotechnol.* **31** (1981) 44 – 54.
82. J.-Y. Wu, *AIChE J.* **44** (1998) 474 – 476.
83. ESSO Research Co., US 3 623 842, 1971 (H. A. Deans).
84. J. F. Tomich, R. L. Dalton, H. A. Deans, L. K. Shallenberger, *J. Pet. Technol.* (1973) 211 – 218.
85. C.-Y. Jeng, S. H. Langer, *J. Chromatogr. Sci.* **27** (1989) 549 – 552.
86. K. F. Scott, C. S. G. Phillips, *J. Chromatogr. Sci.* **21** (1983) 125 – 131.
87. J. E. Patton, S. H. Langer, *Anal. Chem.* **42** (1970) 1449 – 1450.
88. W. R. Melander, H.-J. Lin, J. Jacobsen, C. Horvath, *J. Phys. Chem.* **88** (1984) 4527 – 4536.
89. M. Moriyasu, A. Kato, Y. Hashimoto, *J. Chromatogr.* **400** (1987) 143 – 148.
90. J. Jacobsen, W. Melander, G. Vaisnys, C. Horvath, *J. Phys. Chem.* **88** (1984) 4536 – 4542.
91. M. Lebl, V. Gut, *J. Chromatogr.* **260** (1983) 478 – 482.
92. M. Moriyasu, A. Kato, Y. Hashimoto, *Chem. Lett.* (1984) 1181 – 1184.
93. J. Coca, M. Bravo, E. Abascal, G. Adrio, *Chromatographia* **28** (1989) 300 – 302.
94. S. H. Langer, H. R. Melton, T. D. Griffith, J. Coca, *J. Chromatogr.* **122** (1976) 487 – 503.

Chromatography → Basic Principles of Chromatography

Custom Synthesis and Custom Chemicals

Enrico T. Polastro, Benelux S.A./ N.V., Brussels, Belgium
Arthur D. Little, Benelux S.A./ N.V., Brussels, Belgium

1. **Introduction** 1
2. **Genesis** 2
3. **Supply Structure** 6
4. **Customer Base** 7
5. **Contractual Relations** 8
6. **Key Success Factors** 10
7. **Industry Size and Growth** 11
8. **Concluding Remarks** 12
9. **References** 13

1. Introduction

Custom synthesis can be defined as the business comprising those operations associated with molecule building performed by a third party vendor on a service basis on the behalf of a specific customer having typically no legal affiliation with him.

Custom synthesis is best viewed as one of the multiple segments making the outsourcing space. In its broadest sense this space can be defined as covering all services and products provided on a one-to-one basis by a vendor for the specific use of a given end user. Within this frame, the latter, rather than performing certain activities in-house through his own capabilities and assets, delegates these to a third party.

Although defined as covering only molecule building activities, custom synthesis represents in itself a very complex field made by innumerable segments and subsegments, as illustrated in Figure 1.

These reflect the multitude of situations and variants that can be observed including:

- The type of input supplied, either a service such as the development of an optimized process – technology package ready to be transferred or a product that can be both an intermediate or the final compound.
- The position in the life cycle of the customer end product covered by the custom synthesis relation being considered, namely early and later phase development, launch, maturity or sunset (at peak sales or in their final decline stages). The position of the product in its life cycle has obviously implications in terms of elements such as quantities supplied as well as customer expectations vested in his vendors.
- End-user industry served, e.g., agrochemical, personal care, pharma, electronic chemicals. Molecular structures, volumes as well as regulatory and vendor qualifications hurdles vary widely depending on the application considered.

In custom synthesis both the group of customers and that of vendors is very broad reflecting the extreme fragmentation of this field and the number of situation encountered. Similarly, contractual conditions associated with custom synthesis range widely depending on the scope of the vendor – customer agreement. As an example the agreement can be either restricted to the conversion on a cost-plus basis of a given raw material supplied by the customer or may cover the development by the vendor of the technology – processes required to produce cost effectively the end molecule required by the customer.

It is therefore no surprise that given this extreme fragmentation the precise boundaries and size of the custom synthesis space are ill defined — the industry turnover ranging depending on estimates between more than \$ 50×10^9 to just \$ $6 – 7 \times 10^9$ when limiting its scope to the life science outlet.

Although definitions and boundaries differ depending on the source considered, some traits clearly differentiate custom synthesis from other segments of the chemicals industry. These include:

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.c08_c01.pub2

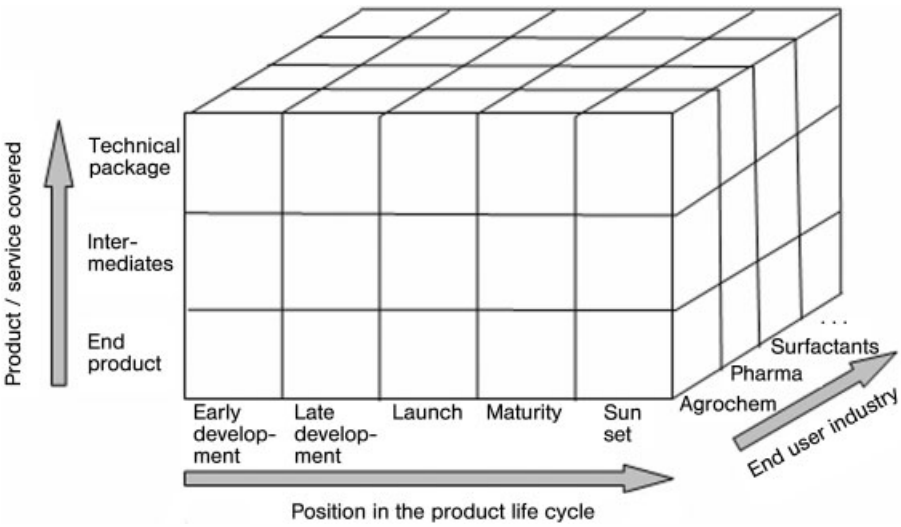

Figure 1. Subsegments of custom synthesis

- The one-to-one relation existing between the vendor and his customer. The first supplying the product or the service on an exclusive basis to the end user who typically is either the originator or owns the full rights on the chemical compound.
- The service type dynamics of the business, the vendor engaging in the supply of the compound at the specific request of a given customer. This means that the vendor is entirely dependent on his customer and has either no or limited control on the ultimate fate of the business should the product be withdrawn from the market or should the customer elect to change of supply source.
- The limited transparency of the business. Customers typically insist to maintain dealings with their vendors confidential, probably as a tactic to increase their leverage.

Although often associated with just fine chemicals and/or the life science industry, custom synthesis covers all segments of the chemical industry and of its outlets and those of its users. Custom synthesis agreements can be observed both for large volume commodity chemicals supplied in tonnages of hundred or thousand of tons and for substances produced on the microgram scale, as an example screening libraries going to the pharma and agrochemical industry. The end-user base for custom synthesis is similarly extremely broad. It comprises both integrated end users having own chemical production facilities electing to outsource part of their requirements, as an example to avoid in-house investments as well as virtual companies having no option but to rely on a third party supplier for their molecular requirements.

The vendor slate is equally extremely broad, in total several thousand companies supplying custom-made chemicals. These include both large diversified chemical companies having developed a base in custom synthesis as a strategy to leverage their technology base and/or to load assets as well as players for whom custom synthesis is their main business. The number of players in custom synthesis has increased since 2000. As an example China and India have become increasingly important forces in custom chemicals, particularly in some product groups, their entry having contributed to exacerbate competitive intensity levels in depressing thereby profit margins in the industry.

2. Genesis

The outsourcing phenomenon dates back from the 1950s. It is finding its roots in the automotive industry having gradually abandoned the full vertical integration model epitomized by the famous “Ford Rivière Rouge” site. On the Rivière Rouge site the Ford Motor Company had built a fully integrated production system able to transform iron ore and coal into complete automo-

biles! However, such an integrated model has been gradually superseded by rather delegating to external suppliers the production and later also the development of selected components or operations of an increasing number of tasks.

The drivers behind the decision to outsource are multiple, including:

- Focus on areas of core competence also often referred to as those operations and processes allowing to effectively create value and providing scope for competitive edge — manufacturing being rarely considered as one of these — contrary to product design, R& D, or marketing.
- The new metrics adopted by management following increasing scrutiny by investors and financial analysts pushing for a more effective use of capital. Within this frame options to avoid in-house investments in fixed assets such as production plants are more and more embraced, outsourcing to a third party being one of these. This drive towards reducing the asset base has prompted some companies to divest part if not their entire manufacturing network.
- The strategy of companies to avoid building up their own fixed cost base — rather electing to outsource some operations. Within this frame outsourcing is viewed as a convenient tool to shift costs to an external entity, allowing to achieve enhanced financial flexibility levels.
- The desire to tap the best available resources when and if required, avoiding building an own resource base that would be difficult to maintain fully loaded and alert.
- Advances in information technology (IT) and communication systems combined with more sophisticated management tools allowing to enhance the transparency of internal operations including in terms of activity costs, enabling also to compare these with alternative options such as outsourcing.
- The emergence of viable alternatives to in-house operations, several vendors focusing their activities around outsourcing services making these their core business. An example is Lonza (Basel, Switzerland) in fine chemicals molecule building.

Parallel to this, a number of new entrants applying innovative business models calling for extensively outsourcing most of their operations have increasingly started to challenge the dominance of traditional fully vertically integrated companies, forcing these to reassess their mode of operations including in terms of sourcing. Examples in the pharmaceutical sector include most emerging pharma players — also often referred to as "biotech" companies — whose focus is on discovery and development.

Similarly to most other segments of the outsourcing space the origins of custom synthesis are difficult to pinpoint to any specific event. The first indications and reports on custom synthesis related activities date from the 1960s and 1970s. At that time several end users — including some large pharmaceutical companies — saw the demand for their products surge, outstripping their own synthesis capacity. Parallel to this, the paradigm calling for backward integration among several end user outlets for chemicals such as pharma, agrochemicals, flavors and fragrances,. . .started to become increasingly challenged. This has led several chemical users to rely on third party vendors not just for "standard" catalogue items but also on products made for their exclusive use under the frame of custom synthesis supply contracts.

One of the first documented examples of custom synthesis appears to be represented in the late 1970s — early 1980s by the supply of cimetidine intermediates to SmithKline & French (Philadelphia, USA). At that time the sales volumes for cimetidine, the first antiulcer drug having truly curative effects allowing to avoid surgery, literally exploded, outstripping the most widely optimistic assumptions. Facing the prospect of losing sales volumes due to short product availability because of chemical synthesis bottlenecks, SmithKline & French contacted various fine chemicals vendors including Fine Organics (Seal Sands, UK), Heico (Philadelphia, USA) and Lonza (Basel, Switzerland) asking them to supply on an exclusive basis selected cimetidine precursors such as the imidazole or the cysteamine intermediate. The overall concept and business model applied was simple, the originator providing to the vendor the process, the vendor applying it in his own facilities. This was corresponding to a win-win scheme:

- The vendor taking advantage of the infrastructure on hand.

- The customer avoiding to invest in own capacity whilst securing immediate access to capacity if and when required.

Since then the custom synthesis area has greatly evolved and grown. Landmarks in this development path include:

- The pioneering efforts of some companies like Lonza in developing the concept of custom synthesis positioning themselves as a credible alternative to in-house chemical operations.
- The growing role played by start-up companies in various end user segments such as pharmaceuticals. Although the share of start-ups in total pharmaceutical sales is still modest - not exceeding a few percentage points of the global pharma turnover $> \$\ 500\ 10^9$– a growing number of new molecular entities (NME) are stemming from these companies (see Fig. 2.).
  As these have no own synthesis capacity, they have no choice but to rely on third party suppliers for their chemical requirements.
- The globalization and increasing transparency of chemical markets combined with the mounting financial pressures facing chemical companies and chemicals users. These are more and more expected to compress their fixed cost base and in parallel reducing capital expenditure (capex) outlays. The industry is gradually starting to pay more attention to concepts such as economic value added or ability to ultimately earn acceptable returns on capital invested. Within this frame a growing number of end users are electing to source their chemical requirements from third party vendors rather than continuing to produce in-house. The rationale behind such outsourcing include on the behalf of the customer considerations such as access to key technologies, capex reduction, cost structure optimization as well as time-to-market imperatives.

Compared to the original days, the custom synthesis industry has become increasingly segmented. As an example, in addition to full scale production of synthetic compounds — be these organic or inorganic, it now covers also biological products obtained by either microbial formulation or animal cell culture as well as related development services:

- Mycophenolic acid — an immunomodulator used to prevent rejection used in organ grafts. The compound, originally developed by Syntex — now Roche (Basel, Switzerland) and obtained through microbial fermentation, has been for long produced on a custom basis by Antibioticos (Settimo Torinese, Italy), a company having a long tradition in fermentation.
- Eternacept, the fusion protein making the active ingredient of Enbrel — one of the most innovative drugs for treatment of degenerative arthritis — is obtained by cell culture. The cultivation has been largely sourced out through custom production agreements between the originator Immunex (Seattle, USA) and Boehringer Ingelheim (Ingelheim, Germany). Amgen (Thousand Oaks, USA) and Wyeth, marketing the product have over the years invested in own capacity.

Although traditionally limited to the supply of full scale volumes, a growing segment of the custom synthesis space is represented by activities associated with the supply of products and services relating to molecules still in early or late development. These include the:

- Supply of libraries of compounds developed at the specific request of a given customer in view of their screening for biological activity to identify possible lead development candidates. These libraries can be either "random" or "focused" — namely revolving around selected molecular platforms, quantities involved rarely exceeding the microgram scale.
- Synthesis of analogues to the original lead compound aiming to obtain suitable development candidates. The synthesis and design of these analogues largely revolves around quantitative structure-activity relationship (QSAR) tools. Such capabilities and services are increasingly offered by third party vendors, examples including Evotec-OSI (Cambridge, UK), AMRI (Albany, USA) or ChemBridge (San Diego, USA).
- Supply of small scale lots of either the final chemical or of its related intermediates in view of their testing on the field. Vendors like ChemShop (Tilburg, Netherlands), Onyx (Newcastle, UK) or Ricerca (Ohio, USA) having made this area their prime focus.

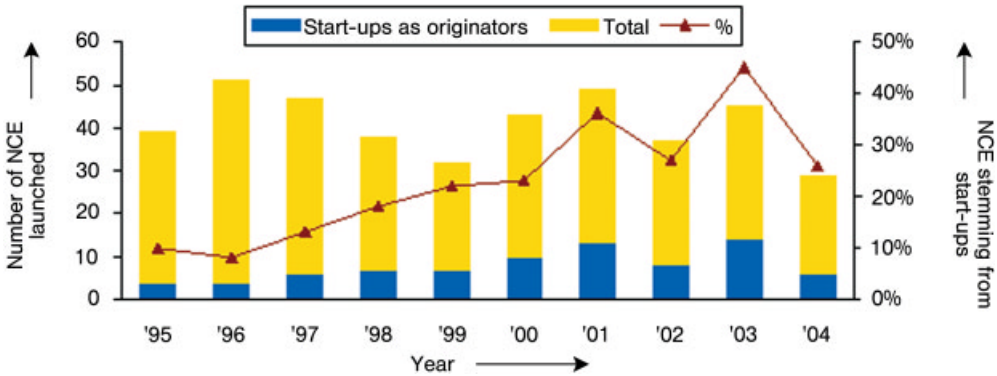

Figure 2. Share of NME by type of originator

- Provision of development services associated as an example to the elaboration of the chemistry, manufacturing, and control (CMC) file, the technical package for compounds still in the development pipeline as well as the design of second generation processes for chemical substances already on the market. The latter corresponds to the request of a customer to a third party vendor to develop a new process allowing to either reduce overall production costs or extend the period of effective patent exclusivity leveraging process related intellectual property (IP).

Another important factor in the overall evolution of the custom synthesis space is the emergence of vendors providing a full range of outsourcing services to the customer — sometimes going beyond molecule building. Examples include in pharmaceuticals, vendors in India such as Matrix (Hyderabad) and Suven (Hyderabad) aiming to become full fledged pharma Custom Research and Manufacturing (CRAM) partners providing to their pharma customers services covering the entire spectrum of molecule building activities — from microgram up to multiton scale — as well as dosage form manufacturing.

In the 1990s industry analysts predicted growing polarization between at one extreme virtual companies entirely outsourcing their chemicals operations and at the other end service providers having no own products but rather supplying on a service basis chemicals to the other cluster of companies. The model postulated that chemical operations are best left to specialized vendors for whom chemical production is their core business — enabling them to achieve better economics through a combination of experience as well as enhanced capital efficiency. Based on these postulates there have been predicaments from the financial community that several chemical companies and chemical users would be better off by entirely farming up their molecule building activities, divesting or spinning off their chemical assets rather relying on specialized third party vendors.

In doing so companies would embrace a "virtual" model delegating to selected vendors major parts if not their entire chemical assets and operations. However, barring a handful of exceptions observed mainly in Japan, hardly any player has moved so far as to fully embrace the virtual model. Rather, most chemical companies and chemical users have now adopted the concept of strategic sourcing where all options to secure access to the required chemical input are systematically evaluated on a which is the best basis. Within this frame neither in-house production nor outsourcing–custom production are viewed as ends perse.

To this end various considerations are taken into account in formulating the sourcing decision, including:

- *Philosophy of the customer and his commitment to chemical operations*. Within this frame European companies tend to be less open to outsourcing than their US counterparts, reflecting their historical roots in chemistry.
- *Position of the product in its life cycle*. The drivers behind outsourcing varying widely between mature to aging products where cost represents a key driver and development compounds for whom availability and time-to-market are critical considerations.

- *Type of technologies and capabilities associated with the production of the compound considered*. Depending on the circumstances very special technologies are either insourced given perceived scope for competitive advantage — or outsourced in order to avoid commit investments in own dedicated assets and resources.
- *Resource availability*. Customers almost invariably tend to prefer loading their own assets and maintain their own staff busy rather than outsource operations and engaging in harsh internal restructuring steps. For example, following the wave of mergers observed in the late 1990s among life science companies an overcapacity situation has emerged. This has resulted in setbacks for custom synthesis service providers as customers have tended during lean times to insource.
- *Security of supply*. Several end users still viewing outsourcing of chemical production more risky compared to in-house operations.
- *Overall economics*. Customers almost invariably tend to compare their own variable costs with price offers of third party vendors, these having necessarily to include full costs incurred plus an acceptable markup.
- *Market access considerations*. Some countries mandate, at least for selected product groups such as agrochemicals and pharmaceutical, local content requirements, de-facto imposing to the marketer to rely on third party suppliers located in the country to be penetrated.
- *Tax optimization strategies*. Here companies leverage transfer prices among different countries — the objective being to log the largest possible share of profits in the lowest tax country. Obviously, within this frame outsourcing requires a careful assessment of its possible impact and portability with tax considerations.

3. Supply Structure

In total probably several thousand companies provide custom synthesis related services. These comprise both small family-owned vendors such as Pressure Chemicals (Chicago, USA) or ChemShop in the Netherlands to divisions of diversified chemical groups, e.g., Dow Pharmaceutical Services and Lonza Exclusive Synthesis. Depending on the supplier custom chemicals can be either the sole business or part of a broader business portfolio.

The emergence of companies specializing on supplying on a custom basis chemical related inputs is to be viewed within the broader context of the reshaping of the chemical industry.

This reshaping has been prompted by drivers including:

- Increasing shareholder activism.
- Chronic inability of most chemical companies to recover their cost of capital over a full economic cycle.
- Surging operating hurdles.
- Emergence of new sets of competitors.
- Evolving customer or more broadly key stakeholders expectations.

As a result the chemical industry has undergone a period of major change and consolidation leading also to the emergence of more focused models such as:

- Chemical utilities, also often referred to as commodity chemical companies focusing on large volume products obtained directly out of basic feedstocks — their value proposition revolving around low cost position reflecting economies of scale and superior access to key feedstocks. Examples of such companies include Methanex (Vancouver, Canada), SQM (Santiago, Chile) or Sabic (Ryadh, Saudi Arabia).
- Product and technology specialists centering their activities around selected product trees and/or technology know-how such as fluorination, nitration, or oxo condensations. Their product range is largely focused on forward commodities and value-added intermediates, levers for value creation being represented by distinctive technology capabilities and focus. Examples of such type companies are by Lonza Polymer Intermediates, Degussa Building Blocks or Rhodia Diphenols.
- Application specialists whose emphasis is on formulated products sold on a performance basis. Key success factors among this set of players largely evolve around application know-how and technical services. Examples of such companies include life science groups, coating and functional food additives suppliers.

- Chemical solution providers whose business is to offer to customers an alternative to own chemical production or formulation activities. Rather than compounds supplied on a specification basis, this set of players is providing customized services to their customers. In doing so, they leverage superior operation management and production economics. Examples of players in this field include in molecule building ChemShop and Lonza Exclusive Synthesis, GE-Betz (Allentown, USA) and PPG (Pittsburgh, USA) providing respectively turn-key chemical solutions in industrial water treatment and automotive coatings.

In addition to Western third party vendors, an increasing number of companies located in China, India, as well as other emerging regions such as CIS, are active in the broadly defined custom chemicals space. Their entry corresponds to the increasing globalization of the chemical industry as well as the quest among customers for the most competitive source:

- CIS has emerged as one of the most important sources of custom-made chemical libraries and early phase custom synthesis services. This country having access to a large molecular diversity pool as well as to substantial synthesis talent. Examples of service providers finding their roots in CIS include ChemBridge (San Diego, USA) as well as Chemical Diversity (San Mateo, USA).
- India is more and more considered for the supply of custom-made intermediates. A growing number of chemical companies trying to build a base in custom chemicals, their value proposition revolving around low labor costs, limited labor and environmental hurdles as well as access to an ample chemically trained work force. Examples of such players include Dishman (Ahmedabad), supplying part of the eposartan bulk requirements of Solvay Duphar, Matrix, or Suven.
- China is taking an increasing role as manufacturing hub for the global chemical industry. Its focus has been traditionally on standard–catalogue products, some companies such as Wuxi Pharmaceuticals (Wuxi) having made out of this business their core activity.

By and large the main attractiveness of Chinese and Indian based vendors results from lower perceived unit prices largely reflecting a combination of cheap labor costs and limited capex outlays compared to their Western counterparts.

The supplier slate in custom chemicals is also extremely heterogeneous in terms of size, comprising both large competitors having custom chemicals sales in excess of $ 300×10^6 such as DPP (DSM Pharmaceutical Products, Heerlen, The Netherlands) or Lonza and smaller companies with a turnover of less than $ 10×10^6, possible examples including Ash Stevens or ChemCon (Freiburg, Germany). Most often smaller competitors tend to focus on niches revolving around selected products or technologies, whilst their larger counterparts offer a broader range of services sometimes positioning themselves as full "molecule building" service providers.

4. Customer Base

The group of customers for custom synthesis is extremely broad. It ranges from major multinational companies such as Procter & Gamble (Cincinnati, OH, USA) or DuPont (Wilmington, DE, USA) to start-up companies like Nicox (Nice, France) or Vertex (Waltham, USA).

The custom synthesis requirements and expectations of these players range widely, including:

- Back-up supply agreements where the customer seeks to secure access to a second source to ensure continuous access to his chemical input requirements.
- The full outsourcing of chemical input requirements from selected third party vendors, the customer having no option but to rely on external sources for his chemical needs. Examples include Wyeth and Gilead,both in pharmaceuticals.

The type of product and input outsourced also spans over a wide spectrum covering:

- The supply of early phase lots or research quantities, as an example screening libraries.
- Intermediates as well as the final compound ready to undergo the formulation/ physical processing step — not requiring to undergo any further molecule building transformation.

Depending on the type of inputs required customer expectations vested in the custom synthesis vendor vary. Most often low delivered unit prices represent the overriding factor for noncritical items such as mature to aging end products where the emphasis is on cost minimization. On the contrary for products in the development phase timely availability, responsiveness, and flexibility represent the critical success factors.

Similarly, the mode of interaction between customers and vendors varies widely, with three main models emerging. These are represented by:

– The traditional "you-against-me" type of interaction where the customer systematically asks bids for every product outsourced from a large number of vendors. The overall driver being to extract all possible price concessions by skillfully playing one vendor against the other.
– The partnership setup where customers rather enter into preferential dealings with one or two main lead — tier 1 — vendors who act as extensions of the customer own operations. In such a setup the customer delegates entire parts of his chemical development and full scale supply requirements to these vendors — setup comparable to the one embraced by the automotive industry. This implies that the retained vendors are granted the right of first refusal on all (most) new chemical requirements of the customer. In exchange of this they are asked to provide preferential treatment of the customer as well as pledging to continuous cost reduction targets to be shared with the customer. Although widely advertised in the industry, examples of genuine strategic partnership agreements between end users and vendors are relatively rare in custom synthesis.
– "A-la-carte" dealings where the customer defines and selects a short list of e.g. five to ten vendors, prequalifying them in terms of range of capabilities, compatibility with own systems and procedures, as well as ability to effectively interface with own staff. In such a setup the customer asks for each project/product offers from two to three vendors, selecting these on a which is the best basis. This à-la-carte mode of interaction appears to be increasingly embraced by the industry.

5. Contractual Relations

Custom synthesis often represents a business where prices of products and services provided are set on a cost-plus basis, the profit margins eventually achievable depending on:

- The overall competitive position including compared to the customer own in-house capabilities.
- The ability of the vendor to efficiently manage his own cost base.

Within this frame it is important to note that typically the customer has a good understanding of the costs involved to produce the inputs he requires, these insights reflecting either his own experience or the price offers received from other vendors. Levers available to vendors to achieve above average returns include:

- Proprietary technologies or IP position, making them the obliged partner to the customer, Lonza Biotechnology with its patented GS (glutamine synthase) cell culture expression system in custom biopharmaceuticals representing a notable example.
- Short capacity availability restricting thereby the customer degrees of freedom in exerting price leverage.

Various pricing schemes can be observed in the custom synthesis space including:

- Fixed price contracts.
- Cost-plus, where the vendor is granted a preagreed markup over and above his own costs.
- Profit sharing, where the vendor receives a fraction of the sales (profit) associated with the end product he is dealing with.

The duration so as the terms of the agreement can vary widely, ranging from yearly (monthly) agreements to multiyears, e.g., three to five years agreements. Depending on the contract the customer can provide for just indicative volume targets, e.g., 80 to 120% of a certain predefined nonbinding target, or settle for firm take-or-pay provisions where the vendor is guaranteed a certain revenue stream irrespective of the quantities of inputs actually supplied. Examples of such take-or-pay agreements include the reservation

fees paid by pharma customers to secure access to biopharmaceutical production capacity following the aftermath of the Etarnecept – Embrel case where supply had been inadequate to meet demand.

Very often a key pricing determinant in custom chemicals is represented by the unit labor and capacity costs incurred by the vendor. Either implicitly or explicitly these are used in the pricing offer provided by the vendor to the customer. In development work a key driver of pricing is the so-called full-time-equivalent (FTE), defined as the price at which the vendor is offering a unit of time (e.g., one hour, one day, a week or a year) for a staff member. Rates applied vary widely. As an example in 2005 FTE observed for process development work ranged from $ 50 000 per person and year to over $ 250 000 per person and year depending on:

- The location of the vendor. The lowest rates tending to be applied by vendors located in China or India whereas North American based companies are among the highest.
- Reputation of the vendor and his track record.
- Range of services included. Some vendors excluding from FTE quotations the supply of raw materials/ reagents, whilst such costs are included in the offers of other suppliers.
- Qualification of the personnel. Some vendors applying identical rates for all their staff contrary to other who differentiate between the rates applied for a Ph.D. and staff without academic degrees.

Similarly for pilot- and full-scale production a key parameter is represented by the equipment hours — namely the cost per unit of time associated with operating a given piece of equipment. Here too the pricing offer of the vendor can be either fully transparent explicitly reporting the estimated hours of utilization for each piece of equipment multiplied by the associated unit costs or rather "flat-inclusive" providing no detail.

In terms of contractual conditions the customer typically insists with his custom chemicals vendors for:

- Full confidentiality, the supply relation being rarely advertised as the customer prefers to keep it secret. The reasons behind such secrecy are multiple, being associated mainly with fears of IP and data leakage. Another reason often indicated relates also with the desire of the customer to keep the supply structure opaque in order to enhance his bargaining position vis-à-vis vendors.
- Complete control and ownership on all IP that may be developed by the vendor as part of the custom chemical supply relationship. In this respect the customer typically demands for being granted automatic licensing right on all know-how and inventions that may be brought by the vendor. Also, most often the customer will request his vendors to refrain from using this know-how for products and/or accounts that may be in direct or indirect competition with him.
- Share all cost savings that may result from learning curve effects and/or improvements brought by the vendor.
- Diligently apply the processes specified in the custom synthesis contract — any change having to be properly documented, reported and approved by the customer before implementation. The adherence to such standard operating procedures (SOP) is particularly critical for chemicals going to applications characterized either by exacting specifications (e.g., electronic chemicals) or extensive regulatory hurdles (e.g., pharmaceutical fine chemicals).

Depending on the contractual agreement, the raw materials can be either supplied by the customer to the vendor or sourced directly by the vendor following the specification set of the customer.

The duration of the contract is also extremely variable, ranging from a few months to multiyear (e.g., five to seven) agreements. Shorter contracts tend to be typical for relatively simple processes that can be easily transferred from one source to the other, as for banal chemical transformations. For more complex processes or where the production site needs to be inspected by the regulatory authorities longer contracts are typically the rule.

Depending on the product, associated volumes as well as on the specific customer the number of custom synthesis vendors retained for a given compound differs. Often small volume products are outsourced to just one vendor, security of supply being managed by maintaining adequate back-up inventories calculated based

on the lead time for developing an alternative source. Multiple sources are established for large volume products as well as molecules generating large sales. For these a lead source providing for, e.g., 70 – 80 % of total volumes is complemented by a back-up vendor — a setup allowing for optimal flexibility and supply reliability.

The custom synthesis contract can cover either the target compound or related intermediates, the customer performing in-house the last synthesis step. The latter setup is typical of several pharmaceutical companies who most often keep in-house these parts of the synthesis process described in the drug master file (DMF) for both quality and tax optimization considerations. As an example, a leading US pharmaceutical company has contracted three vendors to supply the intermediates required for one of its top selling substances, the last couple of synthesis steps being performed in the company's Puerto Rico plant. Such a setup is believed to provide for:

- Optimal efficiency — extracting the best possible price concessions from the three vendors, systematically putting these in direct competition.
- Outstanding capital efficiency — the investment in own synthesis capacity being limited to a bare minimum.
- Superior tax efficiency — performing the last synthesis steps in a tax heaven like Puerto Rico allowing to capture substantial tax concessions.

6. Key Success Factors

Expectations vested by customers in their custom chemicals vendors vary depending on the:

- End product application, itself a key determinant of:
  - Purity requirements.
  - Ultimate impact of the chemical input on the end user price structure and hence at least indirectly scope for value pricing available to the custom chemicals vendor.
  - Regulatory hurdles associated with changes in the production process or in the production site — these being the most stringent in the pharmaceutical outlet.
  - Ultimate tonnages achievable.
- Complexity of the product in terms of elements such as length of the synthesis, type of technologies involved in the production process as well as possible technical hurdles.
- Position of the product in its life cycle.

The most important considerations applied by customers in the selection of vendors include:

- *Competitive unit costs*. These are particularly critical for products in their maturity phase going for commodity applications characterized by intense price competition. Examples of such products include intermediates and actives for the agrochemical industry or for personal care applications. In contrast unit costs are of limited concern for products in their development phase and/or characterized by rich margins such as development pharmaceuticals where timely supply availability rather than price is the key.
- *Supply reliability*. Most often custom chemicals representing for the customer key inputs in view of further processing, penalties associated with supply chain disruption being almost invariably substantial.
- *Service*. The customer expects from his custom chemicals vendors the ability and willingness to meet requests, even if these go beyond what has been agreed in the original contractual arrangements. Examples of such requests may include the supply of odd lots, providing at short notice volumes going beyond what has been indicated in the rolling forecasts, maintaining a buffer stock or accept to withheld an invoice.
- *Technical capabilities*. Customers tending to select vendors on the basis of their (perceived) skills.
- *Responsiveness*. The lead time required by the vendor to respond to the customer requests be these in terms of returning calls or ramp up full-scale production.

These expectations can be translated in terms of key success factors that custom chemicals vendors have to meet to effectively compete. These include:

- Clear positioning, i.e., clearly targeting selected customer groups offering them a credible range of capabilities and services
- Impeccable operations management.

- A service-oriented organization allowing for simple and direct interfaces with the customer. Within this frame a critical element is represented by ensuring a careful alignment and involvement of the entire vendor organization to the customer needs and requirements.
- Standing power, hinging on a combination of factors such as ownership structure, financial strength, environmental situation, etc.
- Transparent communication channels — the customer expecting to be informed on all developments that may have an impact on the products he is outsourcing. These include a change in the process, a new source of raw materials, technology advances that may impact yields. Interesting to note, whilst the customer expects full transparency from his vendors, these are often treated as dispensable resources rather than as genuine business partners!

The role of size as a key success factor in custom chemicals represents a highly controversial topic. At least until recently some industry analysts have been considering size as important, several companies openly stating that sales of at least $ 500 $\times 10^6$ would be needed to continue to effectively compete. Such views were based on the assumption that through a larger size the vendor would be able to increase his standing vis-à-vis customers reducing thereby their ability to exert leverage. Similarly, size has been associated with ability to afford a larger range of resources as well as to better diversify the product portfolio, ensuring thereby more resilience against business vagaries.

Another school of thought is rather viewing size as an obstacle, size going in parallel with increasing organizational complexity with as a corollary loss of flexibility, lower responsiveness and higher fixed costs.

Empirical evidence suggests that an optimum probably exists in terms of size in custom synthesis. At least in Western countries this optimum is in the range of $ 100 to 200 $ $\times 10^6$. Below such as turnover vendors are typically hard pressed to effectively cover their fixed costs and to assemble a well diversified product portfolio. Passed the $ 200 $\times 10^6$ threshold, efficiency and responsiveness tend to decrease often nullifying possible benefits resulting from a larger scale.

7. Industry Size and Growth

No reliable data exist on the size of the custom chemicals space — estimates ranging from less than $ 10 $\times 10^9$ to more than $ 50 $\times 10^9$. This wide range reflects the extreme fragmentation and ill-defined boundaries of custom chemicals whose scope spans across all segments of the chemical industry.

Even, at the individual segment level, such as for example custom-made microbial derived biopharmaceuticals or custom-made libraries of screening compounds for the agrochemical or pharmaceutical industry, it is difficult to precisely estimate the size of the demand. Only orders of magnitude typically exist. This situation originates from the dynamics intrinsic to the custom synthesis business where:

- Transparency is limited, the type of product outsourced so as the source retained being typically kept confidential.
- Only a fraction of the customer product requirements are often outsourced to third party vendors.
- Unit prices paid to the vendor and volumes outsourced are rarely publicly available.

As a result market estimates are most often based on a number of assumptions — some extremely difficult to validate — including:

- Value of the chemical as a percent of the end product sales price. As an example a frequently applied rule of thumb is to put the value of the drug substance and of the agrochemical active at 6 – 8 % and 20 – 25 % of the end product price, respectively. The robustness of such estimates is, however, all but established.
- Share of outsourcing as a percentage of total requirements. Here too only rough estimates are available at least at the consolidated level.

Thus, given the number of hypothesis and assumptions only rough estimates can be provided. As an example the total custom biopharmaceutical market in 2004 is tentatively estimated between $ 1000 $\times 10^6$ and $ 1700 $\times 10^6$ (Arthur D. Little estimates) (High Tech Business Decisions consulting). This broad range reflects at least partly different definitions, the lower estimate

excluding products supplied under the frame of licensing agreements.

Similarly, the size of the custom chemical demand for applications such as detergent components, e.g., optical brighteners or bleaching agents, is difficult to estimate as some supply agreements call for the customer to provide for the raw materials. This situation is in sharp contrast with constructs applied by other companies relying on the vendor to source on his own the required inputs — clearly providing a major source of distortion!

Growth in custom chemicals is equally difficult to precisely assess, growth also often differing depending when measured in volume or value terms. Overall growth of custom chemicals is projected to outpace the corresponding chemical demand and this across most segments. This situation reflects the trend towards increasing outsourcing of chemical operations noted among several end user groups. The drivers behind this trend are multiple including, as explained in more detail in Chapter 2:

- Strategies of end users to focus on key competences and activities truly contributing to create value.
- Attempts to reduce in own capacity and to avoid inflating the fixed cost base.
- Emergence of reliable set of suppliers.

These factors are expected to favor a steady growth of outsourcing. However, it is important to recognize that an upper boundary is unlikely to be exceeded in terms of share of outsourcing as it is not credible to expect that all users will ever entirely embrace a virtual model of chemical operations where they would totally rely on third party vendors!

Also, whilst the growth outlook for custom chemicals is positive in volume terms, unit prices tend to be increasingly under pressure. This is reflecting:

- Leverage applied by customers to their vendors.
- Intense competition between vendors — by and large the custom chemicals space being characterized by an overcapacity situation.
- Inroads of new sets of players including Chinese and Indian vendors offering their services at highly competitive rates, reflecting both lower labor costs and less sophisticated pricing policies, sometimes failing to take into account the cost of capital!

As a result it is common for unit prices associated with the supply of a given custom chemical to decline of a few percentage points every year. This unit price erosion is at least partly offsetting volume growth, resulting for some segments of the custom chemicals space in flat or below GDP growth rates.

8. Concluding Remarks

Custom synthesis has more to do with a service industry than the supply of products. Probably it is the German language and terminology designating custom synthesis as "Lohnsynthese", namely synthesis-at-a-fee that best captures the essence of the business. The vendor is compensated for his molecule building activities performed at the request of a specific customer having retained his services.

Custom synthesis as a business is characterized by highly distinctive dynamics compared to the supply of standard chemicals. These distinctive traits include:

- A substantial interdependence between the vendor and the customer. The vendor is entirely dependent on the customer for each specific product, this dependence link being typically lesser for the customer.
- An extreme fragmentation reflecting the multitude of products, customers, as well as the range of contractual relations existing between the customer and the vendor. This fragmentation provides rich opportunities for small players to successfully compete in the custom chemical space as the scope and need for economies of scale is all but certain.
- Typical service dynamics, the vendor being retained to provide his services on the basis of his capabilities such as R& D or reactor volumes in view of eventually supplying a chemical related output, tailor-made at the specific request of a given customer.

According to some observers custom chemicals represent a highly attractive space, given:

- Predictions of continuous growth fuelled by expanding outsourcing of molecule building activities by end users such as specialty chemicals companies as these focus on areas of

core competence and strive to reduce capex outlays, rather relying on third party vendors.
- Expectations of improving profitability for custom chemicals vendors driven by accelerating supply structure consolidation as small players are squeezed out of the market and customers increasingly rely on their custom vendors for critical inputs — providing thereby scope for value pricing. This would obviously result in lower rivalry between vendors, that combined with lesser customer bargaining power should ultimately translate in expanded profit margins.

These views are strongly challenged by other analysts pointing:

- The limited scope for economies of scale among vendors — failing to support the thesis calling for an overdue supply structure consolidation. Another element appearing to plead against this consolidation is represented by the growing inroads of Chinese and Indian suppliers suggesting that at least in some segments entry barriers are limited.
- The intrinsic disequilibrium in the customer — vendor balance of power. Within this frame the extreme dependence of the vendor on his customer cannot be overemphasized. The vendor having at his disposal very few levers to control his own fate, at least for each individual product. This dependence can be illustrated by the setbacks suffered in the early 1990s by Hickson (Castleford, UK) producing on a custom basis a detergent component for Unilever (Rotterdam, Netherlands), this product being eventually withdrawn from the market. This had far reaching implications for Hickson who saw its capacity utilization and hence revenue stream dramatically fall overnight.

Most probably reality lies in-between these extreme views. Custom synthesis can be attractive, subject to clearly understanding its dynamics and managing these activities accordingly. Also, most importantly, it is critical for vendors already active in this field or planning to enter it to vest realistic expectations on the ultimate profit and growth potential achievable.

9. References

1. Focus report: Active Pharmaceutical Ingredients / Intermediates, Chemical Market Reporter, (2005) March 14, p. FR2-FR18, ICIS Publication, NY.
2. Focus report: Fine Chemicals / Custom manufacturing, Chemical Market Reporter, (2005) January 17, p. FR2-FR17, ICIS Publication, NY.
3. Perspectives for the European Fine Chemical Industry, Miglierini G., *Chimica Oggi/Chemistry Today* **23** (2005) no. 5, Teknoscienze, Milano.
4. Custom manufacturing/Fine chemicals insight: Fine Chemicals: caught in the perfect storm, *Chemical Market Reporter*, **262** (2002) August 12, p. 29, Schenell Publishing, NY.
5. Successful Custom Manufacturing in a highly competitive environment, speech of Mr. Markus Gemund, SOCMS Luncheon Meeting March 21, 2002.
6. Mixed outlook for custom chemicals, *Chem. Eng. News*, **83** (2005) no. 3, January 17, p. 52–57, ACS, Washington DC.
7. Custom Chemical Manufacturing: non pharma pick up, Gordon Graff, Purchasing, (2005) no. 4 March 3, Reed Business Information, Waltham, MA.
8. Custom Chemical Manufacturing: coping with adversity, Gordon Graff, Purchasing, (2004) no. 4, March 4, Reed Business Information, Waltham, MA.
9. Biopharmaceuticals Contract Manufacturing, High Tech Business Decision Rept., 2005 — abstracted in American Pharmaceutical Outsourcing.
10. India emerges in drug services, *Chem. Eng. News, vol.* **83** (2005) no. 5, January 31, p. 15–17, ACS, Washington DC.
11. Setting what you need from your outsourcing partners, The Purple Book 2004, Supplement, SP2 Magazine, Avakado ltd, Horsham, West Sussex.
12. Fine Chemicals: from West to East, E. Polastro, *Scrip Magazine* (2006) April, p. 42, PJB Publications, London.
13. E. Polastro, Facing new challenges in fine chemicals outsourcing, The Purple Book 2005, Supplement, SP2 Magazine, Avakado ltd, Horsham, West Sussex.

HIV and AIDS Therapeutics

Pierre R. Bonneau, Boehringer Ingelheim (Canada) Ltd., Research & Development, Laval (Québec), Canada
Bruno Simoneau, Boehringer Ingelheim (Canada) Ltd., Research & Development, Laval (Québec), Canada

1. Introduction — HIV and AIDS 1
1.1. HIV Cytopathogenesis and AIDS . . . 1
1.2. Virus Life Cycle 2
1.3. Current Targets of Drug Intervention 2
2. Reverse Transcriptase Inhibitors . . . 3
2.1. Introduction 3
2.2. Nucleoside Analogs 4
2.3. Nucleotide Analogs 7
2.4. Non-Nucleoside Analogs 8
2.5. Compounds in Development 9
3. Protease Inhibitors 10
3.1. Introduction 10
3.2. Hydroxyethylamine Isosteres 10
3.3. Hydroxyethylene Isosteres 13
3.4. Non-Peptidic Inhibitors 16
3.5. Compounds in Development 16
4. Entry Inhibitors 17
4.1. Introduction 17
4.2. Fusion Inbihitor 18
4.3. Compounds in Development 18
5. Emerging Concepts in AIDS Therapy 18
6. Acknowledgement 20
7. References 20

The human immunodeficiency virus (HIV) has been identified as the etiologic agent causing the Acquired Immuno Deficiency Syndrome (AIDS). The currently used agents for the treatment of HIV infection mainly target two important viral enzymes or inhibit viral fusion. The nucleoside analogs prematurely terminate the transcription of the viral RNA into dsDNA by reverse transcriptase. The non-nucleoside inhibitors constitute the second class of reverse transcriptase inhibitors. HIV protease, a homodimeric enzyme, is susceptible to inhibition by peptide-like structures. The process of entry of the HIV virus into the target cell can be divided into an attachment step and a fusion step. While only one inhibitor of the fusion step, Enfuvirtide, has entered the market up to now, several small-molecule inhibitors are currently in advanced development. Important efforts are being directed at the development of vaccines that can protect against HIV infection. Continued efforts are being directed at the discovery of therapies which target other essential features of the virus' life cycle, that is, the viral enzyme integrase, the viral maturation process, and the viral infectivity factor.

1. Introduction — HIV and AIDS

The human immunodeficiency virus (HIV) has been identified as the etiologic agent causing the *Acquired ImmunoDeficiency Syndrome* (AIDS) [1, 2]. In the last twenty-five years, AIDS has become the world's most deadly infectious disease, and is the leading cause of death worldwide among those ages 15–59 (20 million deaths since the pandemic began according to a recent World Health Organization report) [3]. It is estimated that 40 million people worldwide are currently suffering from HIV/AIDS infection [4]. About two-thirds of all people living with HIV are in sub-Saharan Africa, as are more than three-quarters of all women living with HIV [4]. The magnitude of this tragedy warrants the unprecedented search for treatment of this deadly disease.

1.1. HIV Cytopathogenesis and AIDS

There are two distinct types of human retroviruses, HIV-1 and HIV-2. Both are members of the lentivirus group of retroviruses. Differences between HIV-1 and HIV-2 are well documented in terms of transmissibility, pathogenesis, and geographic distribution [5, 6]. HIV-1 is widespread across the world and dominates the

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.c13_c01.pub2

pandemic while HIV-2 only occurs at high rates in Western Africa. While the HIV-1 subtype B strains predominate in Europe and the Americas, other HIV-1 subtypes (such as A, C, and D) constitute the majority in Africa [7]. Other subtypes and recombinant forms also increasingly circulate in different parts of the world (such as clade H and CRF02_AG in Central Africa, CRF01_AE in South East Asia [7] and CRF12_BF in South America [8, 9]. Monitoring the antiviral activity of existing and novel anti-HIV agents against this broad and evolving genomic diversity is an aspect of today's therapy.

AIDS results from infection of components of the human immune system by HIV. The consequence of this event is a severe immunosuppression resulting from destruction of CD4-positive T-helper lymphocytes and macrophages (see also → Blood, → Blood, Chap. 2.1.3). HIV can be chronically expressed at a low level for extended periods (average about 10 years from initial infection to clinical disease). During this prolonged disease-free phase, the CD4 cell-count slowly decreases to a level below which the immune system is no longer effective against opportunistic pathogens, resulting in the emergence of multiple infections, tumors, and, ultimately, death [10–12].

1.2. Virus Life Cycle

The HIV-1 genome consists of ten open reading frames encoding structural, regulatory and accessory proteins (Table 1, Fig. 1) [13–15]. Products of the *gag* gene form the major components of the virion core. The *pol* gene encodes all important enzymes required for DNA synthesis and integration, and processing of viral polyproteins. The *env* gene product is a structural glycoprotein that is important in AIDS pathogenesis. Regulatory proteins (*tat*, *rev* gene products, as well as others shown in Table 1) are required for control of the viral replication cycle. Finally, accessory gene products (*vpu*, *vif*) perform important functions related to virion maturation, release and infectivity.

HIV infection is mediated by the sequential interaction of the *gp*120 glycoprotein on the surface of the virus with the cellular receptor molecule, the CD4 glycoprotein, and then a coreceptor, which are usually the chemokine receptor molecules CCR5 and CXCR4 (Fig. 1). These interactions lead to the activation of the fusogenic transmembrane viral glycoprotein gp41. Following fusion with the host cell membrane and uncoating of the viral particle, viral single-stranded RNA is transcribed into double-stranded DNA (dsDNA) by the reverse transcriptase (RT) enzyme. After transport into the cell nucleus, integration of viral dsDNA by the integrase (IN) enzyme into the cellular chromosomal DNA generates *provirus*. The virus has now permanently infected the cell. Following a latency period, expression of the provirus leads to virus replication. Viral messenger RNAs are spliced and transported into the cytoplasm where viral protein synthesis occurs. In the final stages of replication, *gag* and *gag-pol* precursor proteins and two molecules of viral RNA assemble with envelope protein to form an immature virion. The autocatalytically released HIV protease then cleaves the remaining protein precursors to form a new, mature and infectious viral particle that is released from the cell.

Table 1. HIV-gene functions

Genes	Functions
Structural	
gag	matrix, virion maturation and stability, capsid, virus particle maturation and release
pol	proteolysis of viral polyprotein nucleic acid replication and integration
env	external envelope, receptor binding, virion infectivity
Regulatory	
tat	transcription transactivator increasing viral gene expression
rev	rna binding protein, regulator of protein expression
tev/tnv	undefined
nef	negative factor, virus propagation
vpr	early regulatory protein
Accessory	
vif	viral infectivity factor, cell-free virus transmission, *env* processing
vpu	virus maturation/release

1.3. Current Targets of Drug Intervention

Intense studies on the replication cycle of the virus continue to provide opportunities for therapeutic intervention [16]. So far the currently used agents for the treatment of HIV infection mainly target two important viral enzymes (the

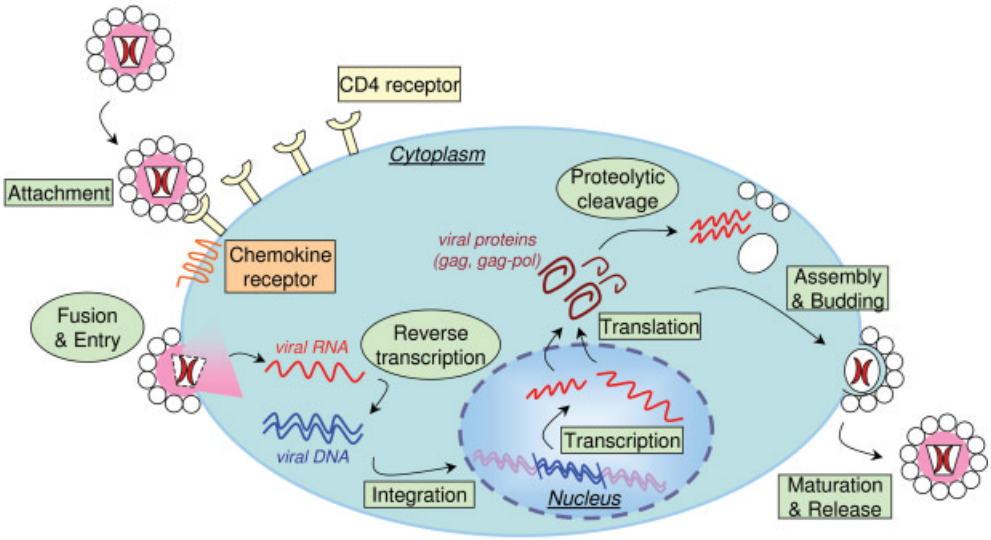

Figure 1. HIV virus life cycle

early stage reverse transcriptase and the later stage protease), or inhibit viral fusion.

To infect a cell, the HIV virus must first fuse its membrane with the cellular membrane. This process is mediated by the viral envelope glycoprotein on the surface of the virus, which is composed of trimers of noncovalently associated heterodimers of gp120 and gp41. Following interaction of gp120 with the cellular receptors (CD4 and CCR5/CXCR4), the N-terminal hydrophobic fusion peptide of gp41 inserts into the cellular membrane and triggers extensive conformational changes resulting in the coming together of the viral and cellular membranes and ultimately fusion. The discovery that inhibition of this process prevented infection led to the successful development of a peptidic inhibitor of fusion.

Reverse transcriptase is unique to retroviruses and since no analogous enzyme activity is present in normal mammalian cells, inhibitors of this enzyme were viewed as particularly attractive therapeutics, with the potential for a large margin of safety [17]. Reverse transcriptase transforms the single-stranded RNA virus genome into dsDNA, prior to integration into the host genome. Firstly, it transcribes a strand of DNA that is complementary to the viral RNA. Secondly, as this process is occurring, reverse transcriptage degrades the original RNA strand (using the ribonuclease H domain of reverse transcriptase), leaving a single stranded DNA that in turns serves as template for the synthesis of a complementary [18] DNA strand, resulting in dsDNA. The first licensed anti-HIV inhibitors were targeted at reverse transcriptase and since then several new compounds with improved properties have increased the list.

HIV protease, another essential protein in the virus life cycle, processes the *gag* and *gag-pol* gene products. Cleavage of these polyproteins by the protease at specific sites generates the proteins and enzymes required for viral maturation [19, 20]. The discovery that inhibition of the protease resulted in the production of non-infectious, immature virions [21, 22] led to the development of several inhibitors of this enzyme [23–26].

2. Reverse Transcriptase Inhibitors

2.1. Introduction

Two classes of reverse transcriptase inhibitors are used in clinical practice. The first identified were the *nucleoside analogs* (commonly referred to as NRTIs), which constitute the backbone of first line therapeutic intervention [27] and most subsequent lines. After conversion to their triphosphates, which are the pharmacologically active species within cells, the nucleoside analogs mimic the natural substrates of the en-

zyme. However, since the analogs lack the 3′-OH group necessary for DNA chain elongation, they prematurely terminate the transcription of the viral RNA into dsDNA. They may also act as competitive inhibitors with respect to the natural nucleoside substrates. More recently tenofovir, a *nucleotide* reverse transcriptase inhibitor, was added to the list of approved chain terminators. These inhibitors typically still have some affinity for endogenous DNA polymerase enzymes, which can lead to limiting side effects.

The *non-nucleoside inhibitors* (commonly referred to as NNRTIs) constitute the second class of reverse transcriptase inhibitors. In contrast to the nucleoside analogs, these drugs bind to a lipophilic site of the reverse transcriptase located close to the active site and reduce the conformational flexibility of the enzyme. Non-nucleoside inhibitors are characterized by extremely high selectivity for reverse transcriptase, and are ineffective even against the closely related HIV-2 reverse transcriptase. These drugs are used in combination with the various nucleoside reverse transcriptase inhibitors in multiple therapy regimens. The use of NNRTIs is however limited by extensive cross-resistance within this class.

2.2. Nucleoside Analogs

Zidovudine [*30516-87-1*], 3′-azido-3′-deoxythymidine, AZT, $C_{10}H_{13}N_5O_4$, M_r 267.2, *mp* 116–118 °C [28, 29], was the first nucleoside inhibitor approved for use in the treatment of AIDS in 1987, and still constitutes an important component of the backbone of combination regimens that include other nucleoside reverse transcriptase inhibitors, nonnucleoside reverse transcriptase inhibitors and protease inhibitors.

Synthesis [28, 29] (see scheme on top of the next column):

Trade names (zidovudine): Retrovir® (GlaxoSmithKline), Combivir® (coformulation tablet of zidovudine and lamivudine, GlaxoSmithKline). Trizivir® (coformulation tablet of zidovudine, lamivudine and abacavir sulfate, GlaxoSmithKline)

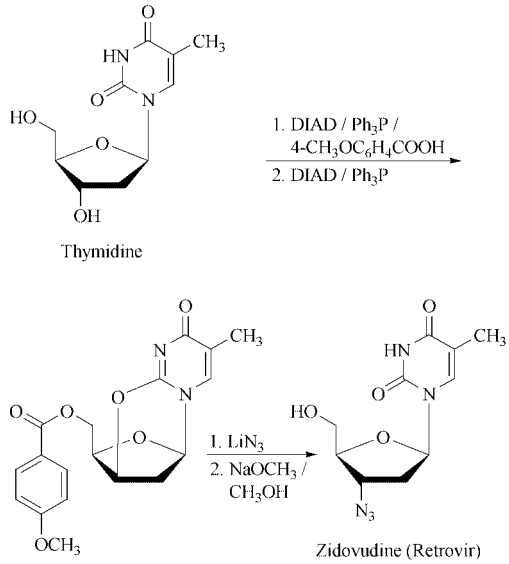

Didanosine [*69655-05-6*], 2′,3′-dideoxyinosine, ddI, $C_{10}H_{12}N_4O_3$, M_r 236.2, *mp* 177–180 °C (several values in the literature).

Synthesis [30]:

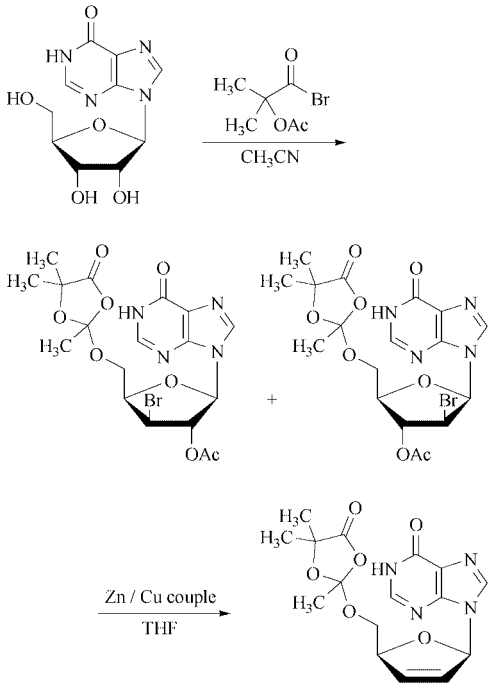

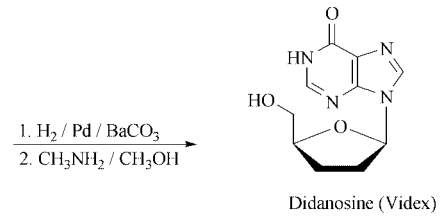

Trade name: Videx® (Bristol-Myers Squibb Co).

Zalcitabine [*7481-89-2*], 2′,3′-dideoxycytidine, ddC, $C_9H_{13}N_3O_3$, M_r 211.2, *mp* 215–217 °C. ddC is no longer a recommended component of therapy and has been withdrawn in some countries.

Synthesis [31]:

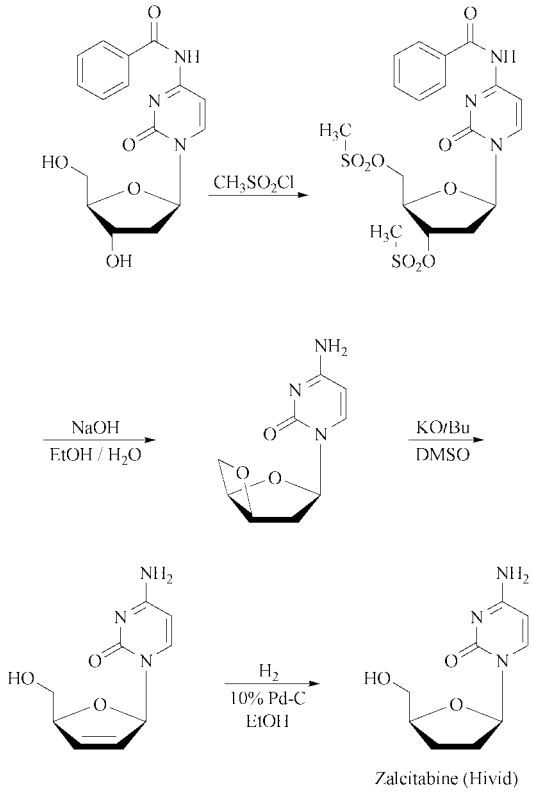

Trade name: Hivid® (Hoffmann-LaRoche).

Stavudine [*3056-17-5*], 2′,3′-didehydro-3′-deoxythymidine, d4T, $C_{10}H_{12}N_2O_4$, M_r 224.2, *mp* 165–166 °C.

Synthesis [32]:

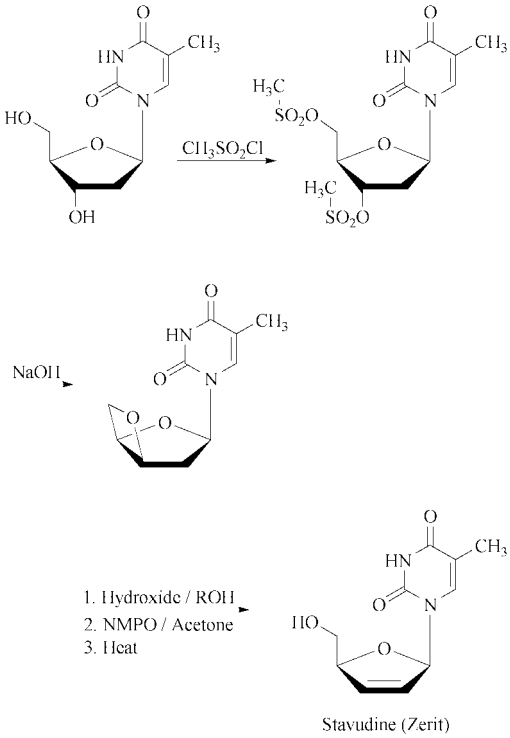

Trade name: Zerit® (Bristol-Myers Squibb Co).

Lamivudine [*134678-17-4*], (2*R*,5*S*)-4-amino-1-[2-(hydroxymethyl)-1,3-oxathiolan-5-yl]-2(1*H*)-pyrimidinone, 3TC, $C_8H_{11}N_3O_3S$, M_r 229.3, *mp* 158–160 °C.

Synthesis [33, 34]:

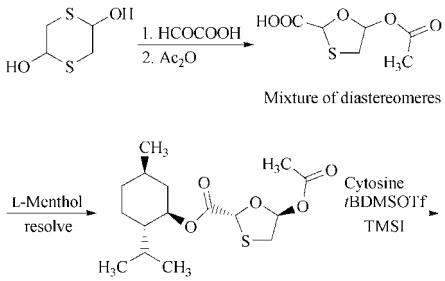

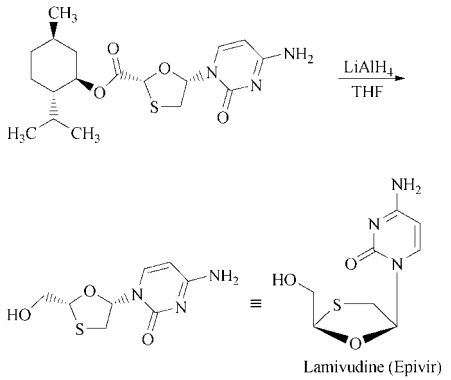

Trade names: Epivir® (GlaxoSmithKline), Combivir® (coformulation tablet of lamivudine and zidovudine, GlaxoSmithKline). Trizivir® (coformulation tablet of zidovudine, lamivudine and abacavir sulfate, GlaxoSmithKline). Epzicom™, Kivexa™ (coformulation tablet of lamivudine and abacavir sulfate, GlaxoSmithKline).

Abacavir [*136470-78-5*], (1*S*,4*R*)-4-[2-amino-6-(cyclopropylamino)-9*H*-purin-9-yl]-2-cyclopentene-1-methanol, $C_{14}H_{18}N_6O$, M_r 286.3, foam; abacavir sulfate [*188062-50-2*], (1*S*,4*R*)-4-[2-amino-6-(cyclopropylamino)-9*H*-purin-9-yl]-2-cyclopentene-1-methanol, sulfate salt (2:1), $(C_{14}H_{18}N_6O)_2H_2SO_4$, M_r 670.76, *mp* 224–225 °C.

Synthesis [35, 36]:

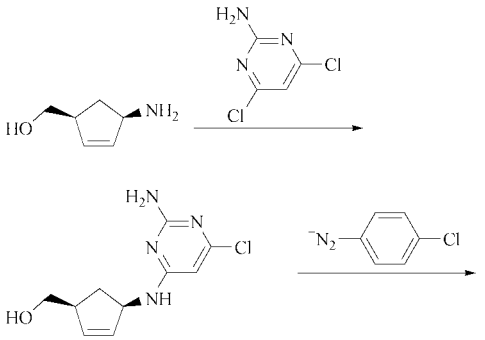

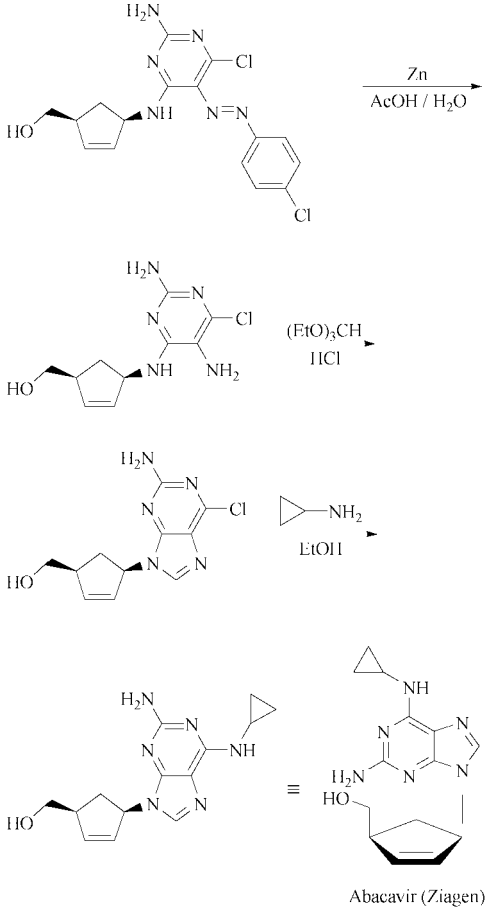

Trade name (abacavir sulfate): Ziagen® (GlaxoSmithKline). Trizivir® (coformulation tablet of zidovudine, lamivudine and abacavir sulfate, GlaxoSmithKline). Epzicom™, Kivexa™ (coformulation tablet of lamivudine and abacavir sulfate, GlaxoSmithKline).

Emtricitabine [*143491-57-0*], 4-amino-5-fluoro-1-[(2*R*,5*S*)-2-(hydroxymethyl)-1,3-oxathiolan-5-yl]pyrimidin-2-(1*H*)-one, FTC; $C_8H_{10}FN_3O_3S$, M_r 247.24, *mp* 166°C (form II) [37].

Synthesis [38, 39]:

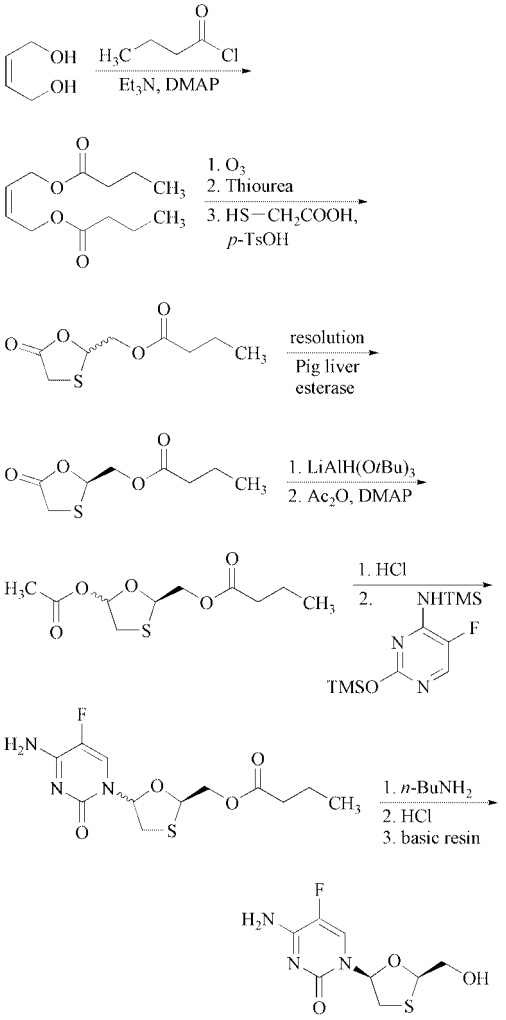

Emtricitabine (Emtriva)

Trade name (emtricitabine): Emtriva® (Gilead Sciences Inc.). Truvada® (co-formulation tablet of emtricitabine and tenofovir disoproxyl fumarate) (Gilead Sciences Inc.)

2.3. Nucleotide Analogs

Tenofovir [*147127-20-6*], [[(1R)-2-(6-amino-9H-purin-9-yl)-1-methylethoxy]methyl]-phosphonic acid; $C_9H_{14}N_5O_4P$ M_r 287.21.

Tenofovir disoproxil fumarate [*202138-50-9*] is a prodrug of tenofovir. Its chemical name is bis{[(isopropoxycarbonyl)oxy]methyl}{[(1R)-2-(6-amino-9H-purin-9-yl)-1-methylethoxy]methyl}phosphonate;fumarate (1:1), $C_{19}H_{30}N_5O_{10}P{\cdot}C_4H_4O_4$ M_r 635.52, *mp* 114.2–114.6°C [40].

Synthesis [40]:

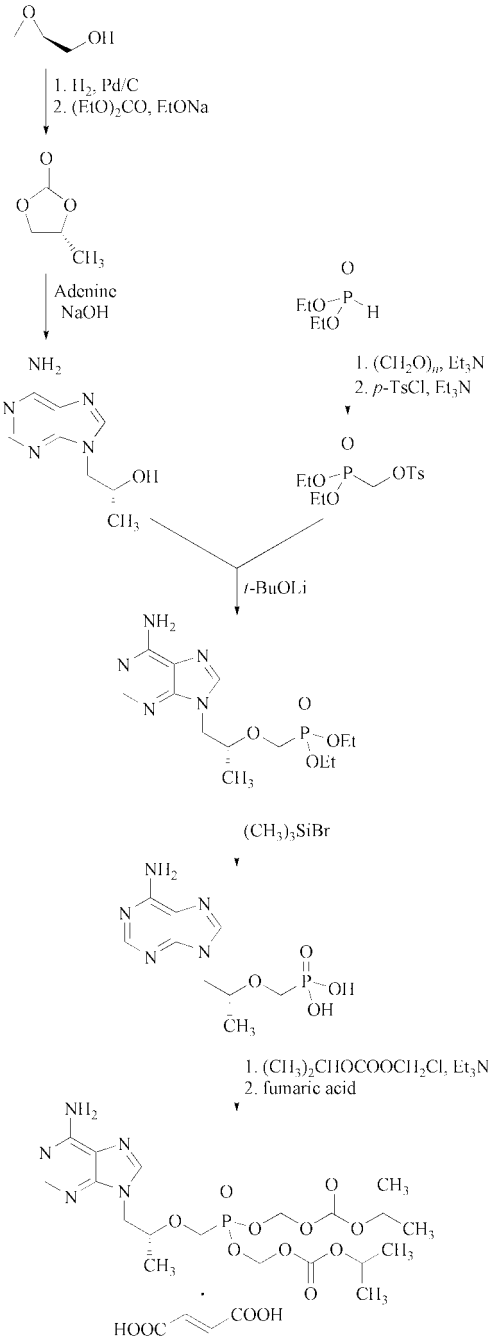

Tenofovir disoproxil fumarate (Viread)

Trade name (tenofovir disoproxil fumarate): Viread® (Gilead Sciences Inc.) Truvada® (co-formulation tablet of emtricitabine and tenofovir disoproxil fumarate) (Gilead Sciences Inc.).

2.4. Non-Nucleoside Analogs

Nevirapine [*129618-40-2*], 11-cyclopropyl-5,11-dihydro-4-methyl-6*H*-dipyrido[3,2-b:2′,3′-e][1,4]diazepin-6-one, $C_{15}H_{14}N_4O$, M_r 266.3, *mp* 247–249 °C, is a non-competitive inhibitor of HIV-1 reverse transcriptase, and was the first of the non-nucleoside reverse transcriptase inhibitor class to receive approval for use in the treatment of AIDS in 1996.

Synthesis [41]:

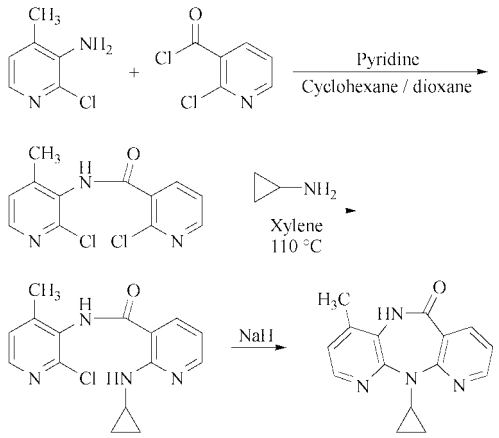

Nevirapine (Viramune)

Trade name: Viramune® (Boehringer Ingelheim).

Delavirdine [*136817-59-9*], 1-[3-[(1-methylethyl)amino]-2-pyridinyl]-4-[[5-[(methylsulfonyl)amino]-1*H*-indol-2-yl]carbonyl]piperazine, $C_{22}H_{28}N_6O_3S$, M_r 456.56, *mp* 226–228 °C; delavirdine methanesulfonate (mesylate) [*147221-93-0*] $C_{22}H_{28}N_6O_3S \cdot CH_4O_3S$, M_r 552.68, *mp* 220–222 °C.

Synthesis [42, 43]:

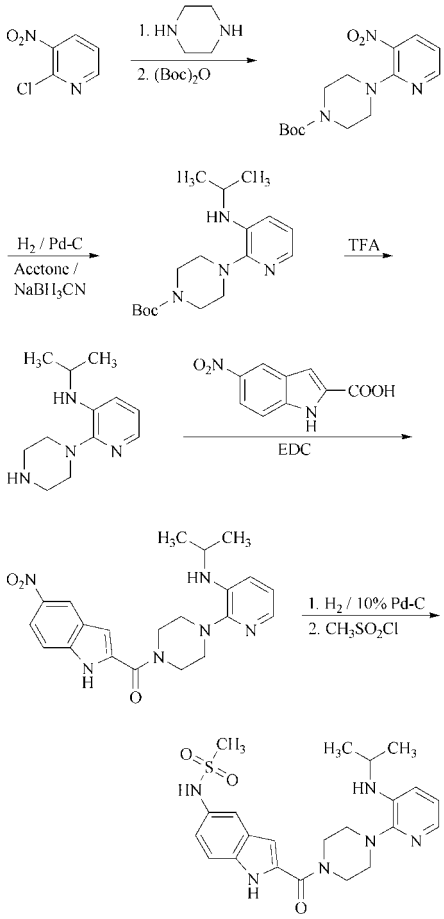

Delavirdine (Rescriptor)

Trade name (delavirdine mesylate): Rescriptor® (Pfizer, registered trademark of Pharmacia & Upjohn Co).

Efavirenz [*154598-52-4*], (4*S*)-6-chloro-4-(cyclopropylethynyl)-1,4-dihydro-4-(trifluoromethyl)-2*H*-3,1-benzoxazin-2-one, $C_{14}H_9ClF_3NO_2$, M_r 315.7, *mp* 139–141 °C.

Synthesis [44]:

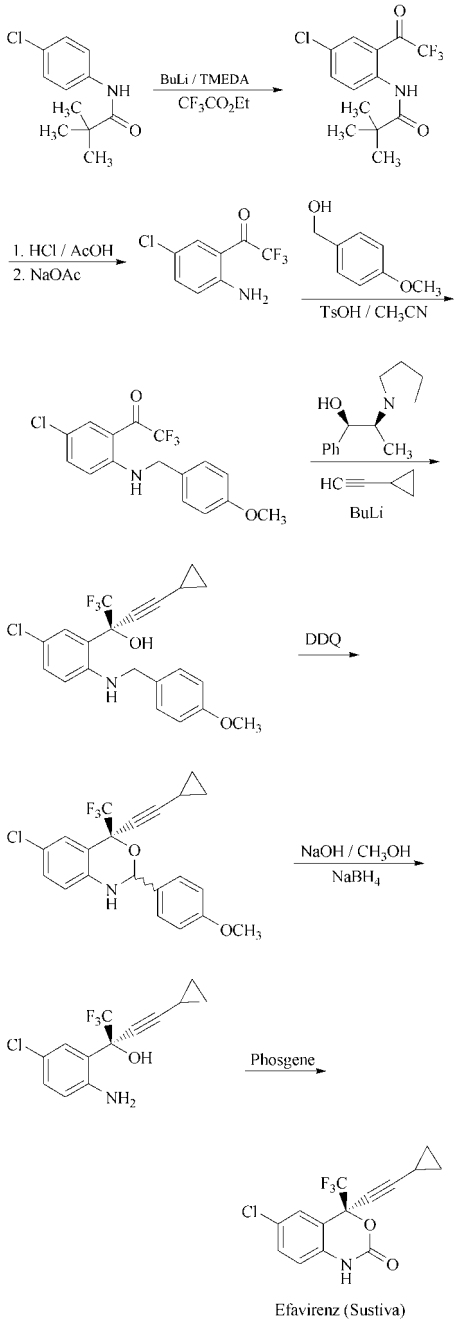

Trade name (efavirenz): Sustiva™ (Bristol-Myers Squibb Co), Stocrin® (Merck & Co.).

2.5. Compounds in Development

Several additional nucleoside and non-nucleoside reverse transcriptase inhibitors are currently in various stages of clinical development. In general, these drug candidates attempt to improve on the various limitations associated with the currently used therapies. Emphasis is on safer and more convenient dosages as well as on the development of new agents with broader antiviral profiles against reverse transcriptase mutations associated with reduced susceptibility to currently approved NRTIs and NNRTIs.

Among the NRTIs in advanced stages of development, one can list: Elvucitabine (L-D4FC, ACH-126443) [*181785-84-2*] by Achillion Pharmaceuticals [45], Dexelvucitibine (D-d4FC, DPC-817) [*134379-77-4*] by Pharmasset [46], Racivir (racemic FTC) [*143491-54-7*] by Pharmasset [47], and AVX-754 (Apricitabine SPD-754, (−)-dOTC) [*160707-69-7*] by Shire/Avexa Limited [48].

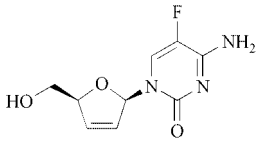

Dexelvucitibine

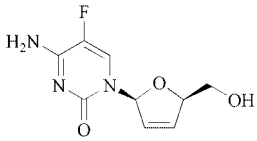

Elvucitabine

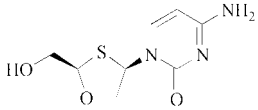

SPD-754

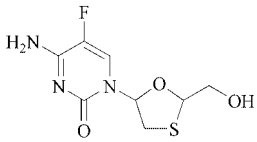

Racivir (racemic FTC)

Among the NNRTIs in advanced stages of development, one can list: Etravirine (TMC-125, R-165335) [*269055-15-4*] [49] and

rilpivirine (TMC-278, R-278474) [*500287-72-9*] by Tibotec-Virco/Johnson& Johnson [50], capravirine (S-1153, AG-1549) [*178979-85-6*] by Shionogi [51], GW-5634 (GW-695634) [*457635-64-2*] a pro-drug of GW-8248 by GlaxoSmithKline [52], and BILR 355 BS [*380378-81-4*] by Boehringer Ingelheim [53, 54]. Capravirine and GW-5634 have recently being discontinued.

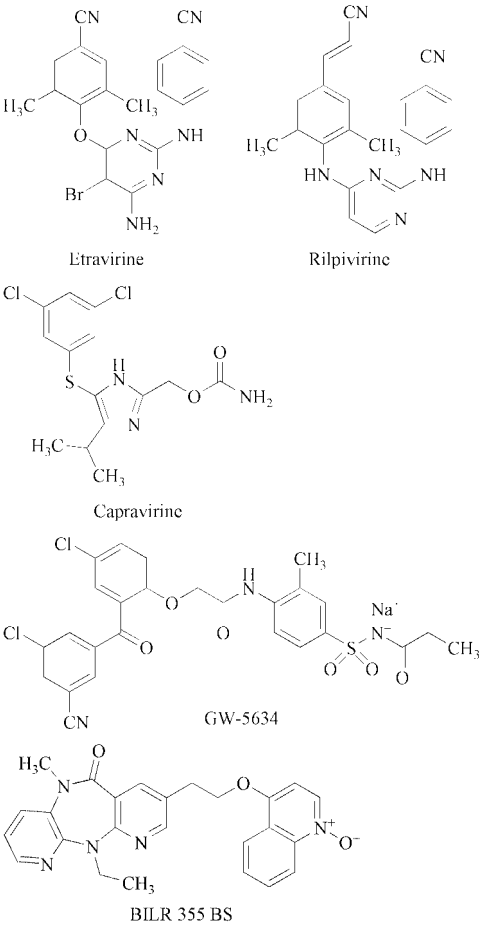

3. Protease Inhibitors

3.1. Introduction

Being a member of the aspartyl protease class of enzymes [26, 55], HIV protease, a homodimeric enzyme, is susceptible to inhibition by peptide-like structures incorporating various isosteres that mimic the transition state for the hydrolysis of an amide bond (Fig. 2) [23–25].

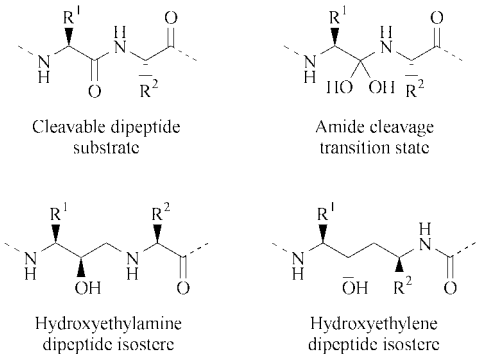

Figure 2. Protease inhibitors mimic the transition state of hydrolysis of an amide bond

Among them, hydroxyethylamine and hydroxyethylene isosteres have led to potent inhibitors of the enzyme. Despite their peptidic character (often resulting in low oral bioavailability and rapid excretion), high molecular masses (575–750) and structural complexities, these classes of compounds effectively reduce viral load in infected patients and provide physicians with additional tools for combating the deadly disease. In 2005, a non-peptidic protease inhibitor (tipranavir, Aptivus®) was approved bringing to nine the number of HIV protease inhibitors on the market. Nowadays, most protease inhibitors are often co-administered with low-dose ritonavir (another HIV protease inhibitor also inhibiting the metabolic CYP 3A4) in order to improve their pharmacokinetic properties and to lower dosage [56, 57].

3.2. Hydroxyethylamine Isosteres

Saquinavir [*127779-20-8*], [3*S*-[2[1*R**(*R**), 2*S**],3 α, 4a β, 8a β]]-N^1[3-[3-[[(1,1-dimethylethyl)amino]carbonyl]octahydro-2(1*H*)-isoquinolinyl]-2-hydroxy-1-(phenylmethyl)propyl]-2-[(2-quinolinylcarbonyl)amino]butanediamide; free base, $C_{38}H_{50}N_6O_5$, M_r 670.85, $[\alpha]_D^{20}$ −55.9 ° (c = 0.5, MeOH); monomethanesulfonate (mesylate) [*149845-06-7*], $C_{38}H_{50}N_6O_5 \cdot CH_4O_3S$, M_r 766.95.

Saquinavir was the first peptidomimetic-based HIV protease inhibitor to be approved in the United States (1995) and Europe (1996) [58–60]. A new soft gel formulation (Fortovase®) was launched in the United States (1997) and provides improved bioavailability.

Synthesis [61]:

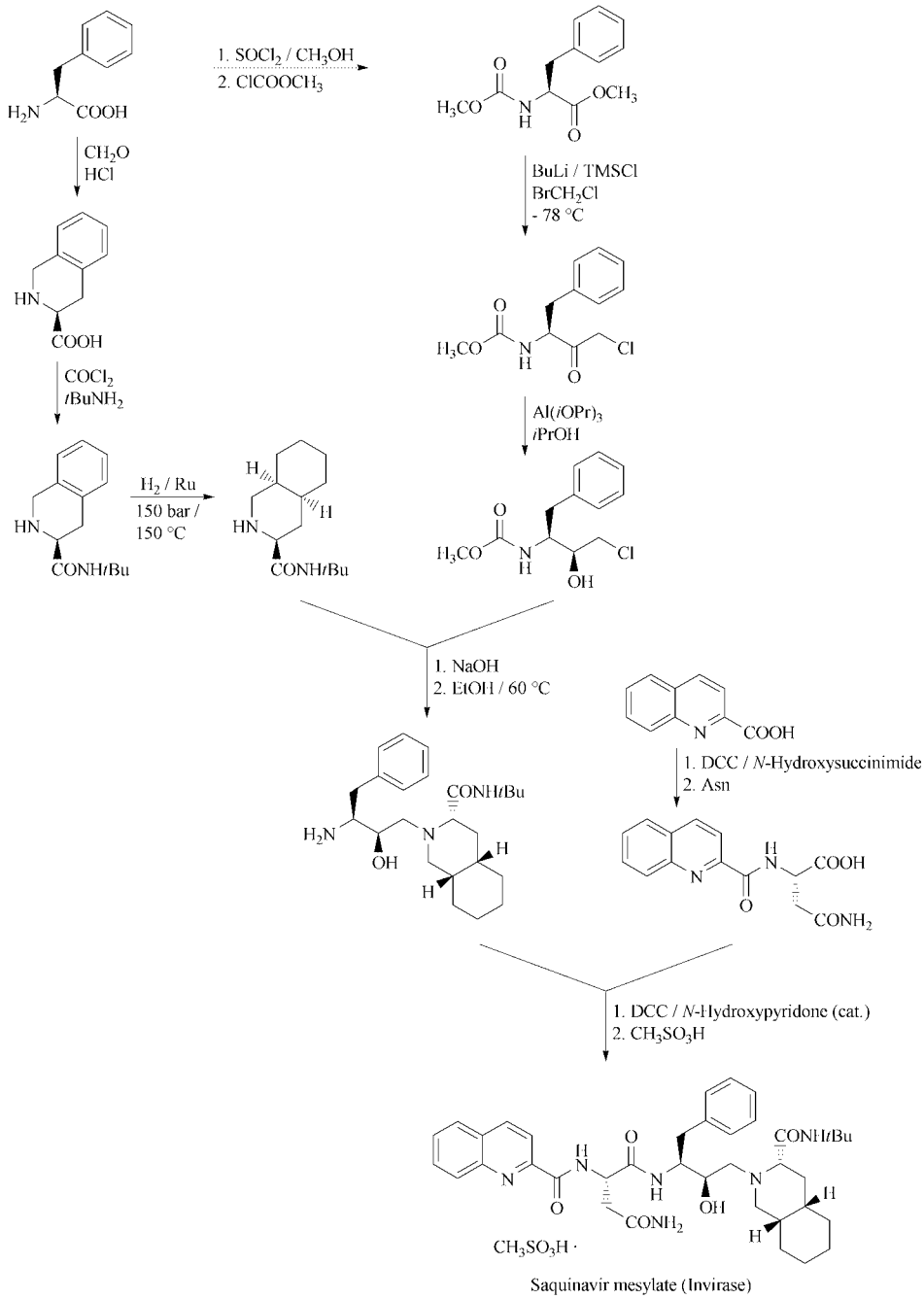

Trade names: Invirase® (saquinavir mesylate, Ro 318959, Hoffmann-La Roche, Switzerland, must be co-administered with ritonavir to improve plasma levels). Fortovase® (soft gel formulation, Ro 318959 003, Hoffmann-La Roche, Switzerland).

Nelfinavir [*159989-64-7*],[3*S*-[2(2*S**,3*S**), 3 α 4a β, 8a β]]-*N*-(1,1-dimethylethyl)decahydro-2-[2-hydroxy-3-[(3-hydroxy-2-methylbenzoyl)amino]-4-(phenylthio)butyl]-3-isoquinolinecarboxamide; free base $C_{32}H_{45}N_3O_4S$, M_r 567.80; monomethanesulfonate (mesylate) [*159989-65-8*], $C_{32}H_{45}N_3O_4S \cdot CH_4O_3S$.

Synthesis [62, 63]:

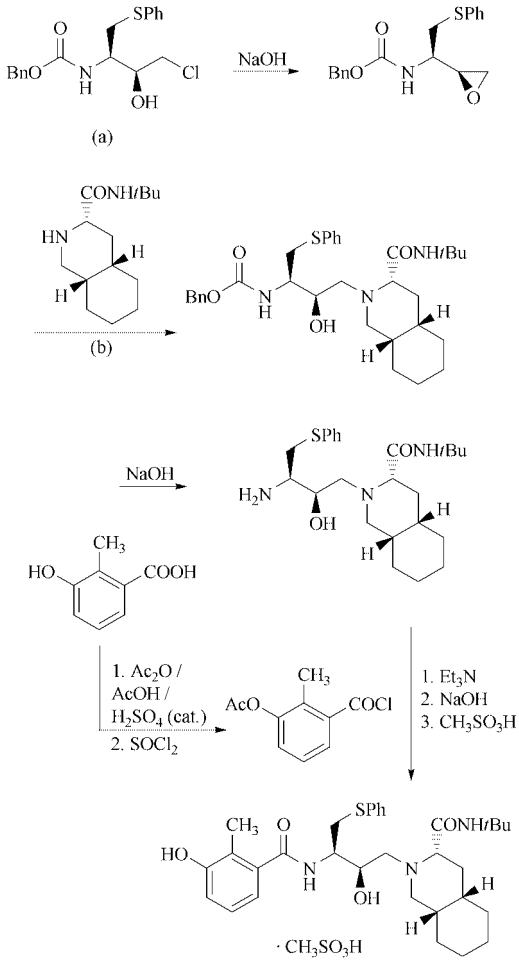

Nelfinavir mesylate (Viracept)

(a: [64], b: [65])

Trade names (nelfinavir mesylate): Viracept® (PfizerAgouron, Japan Tobacco; Hoffmann-La Roche).

Amprenavir [*161814-49-9*], (3*S*)-tetrahydro-3-furyl N-[(1*S*,2*R*)-3-(4-amino-N-isobutylbenzenesulfonamido)-1-benzyl-2-hydroxypropyl]carbamate; free base $C_{25}H_{35}N_3O_6S$, M_r 505.64.

Synthesis [66]:

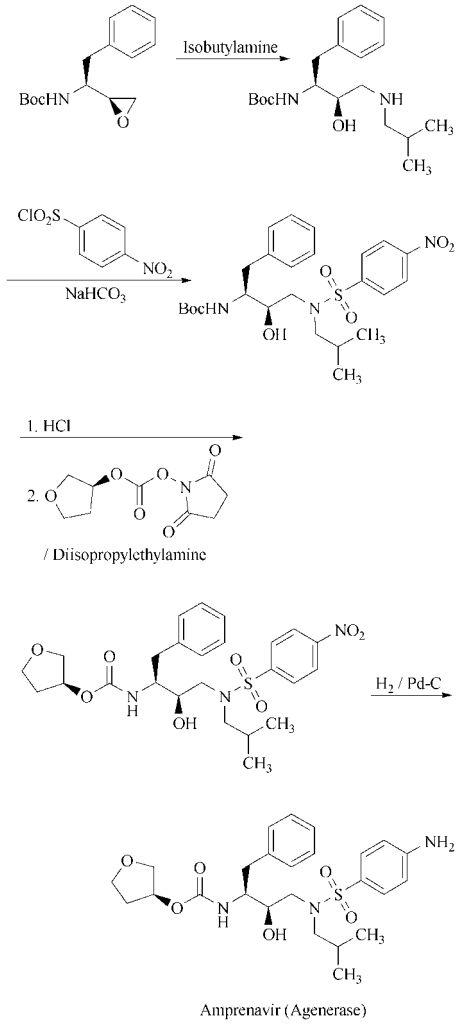

Amprenavir (Agenerase)

Trade name: Agenerase® (GlaxoSmithKline; licensed from Vertex Pharmaceuticals Inc), Prozei® (Kissei Pharmaceutical, Japan; licensed from Vertex Pharmaceuticals Inc.).

Fosamprenavir calcium [*226700-81-8*], (3*S*)-tetrahydrofuran-3-yl (1*S*,2*R*))-3-[[(4-aminophenyl)sulfonyl](isobutyl)amino]-1-benzyl-2-(phosphonooxy)propylcarbamate monocalcium salt; $C_{25}H_{34}CaN_3O_9PS$, M_r 623.7, *mp* 282–284 °C is a prodrug of amprenavir that was approved in 2003 for use in combination with

reverse transcriptase nucleoside / nucleotide analogs without or with ritonavir.

Synthesis [67, 68]:

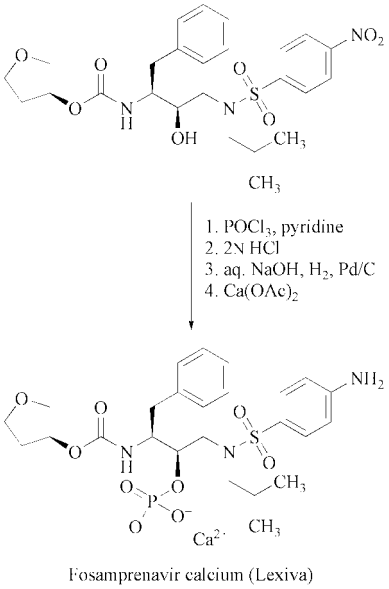

Fosamprenavir calcium (Lexiva)

Trade name: Lexiva®, Telzir™ (GlaxoSmithKline; licensed from Vertex Pharmaceuticals Inc., USA).

3.3. Hydroxyethylene Isosteres

Ritonavir [*155213-67-5*], [5*S*-(5*R**,8*R**,10*R**,11*R**)]-10-hydroxy-2-methyl-5-(1-methylethyl)-1-[2-(1-methylethyl)-4-thiazolyl]-3,6-dioxo-8,11-bis(phenylmethyl)-2,4,7,12-tetraazatridecan-13-oic acid 5-thiazolylmethyl ester; free base $C_{37}H_{48}N_6O_5S_2$, M_r 720.96.

Ritonavir is a peptidomimetic inhibitor of HIV-1 and HIV-2 proteases that was approved in 1996 for use in combination with reverse transcriptase nucleoside analogs. It reversibly and potently inhibits cytochrome P450 enzymes, thus inhibiting not only its own metabolism, but also inhibiting the metabolism of other drugs to which it is combined [69]. For this reason, ritonavir has found utility as a boosting agent when administered at low dose in combination with certain anti-HIV agents to improve their pharmacokinetic properties (such as exposure and half-life) [56, 57].

Synthesis [70–72] (see next page)

Indinavir [*150378-17-9*], [1(1*S*,2*R*),5(*S*)]-2,3,5-trideoxy-*N*-(2,3-dihydro-2-hydroxy-1*H*-inden-1-yl)-5-[2-[[(1,1-dimethylethyl)amino]carbonyl]-4-(3-pyridinylmethyl)-1-piperazinyl]-2-(phenylmethyl)-D-erythropentonamide; free base $C_{36}H_{47}N_5O_4$, M_r 613.80, *mp* 153–154 °C (anhydrous form), $[\alpha]_D^{20}$ + 24.1 ° (*c* = 0.0133, $CHCl_3$); sulfate [*157810-81-6*], $C_{36}H_{47}N_5O_4 \cdot H_2SO_4$, M_r 711.88, crystals from absolute ethanol, softens at 135 °C, *mp* 150–153 °C (decomp.).

Synthesis [73]:

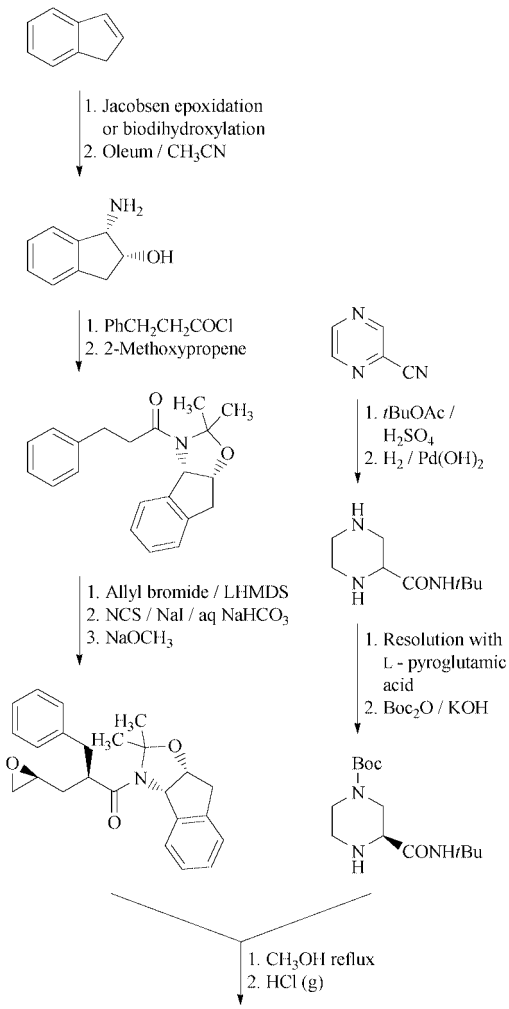

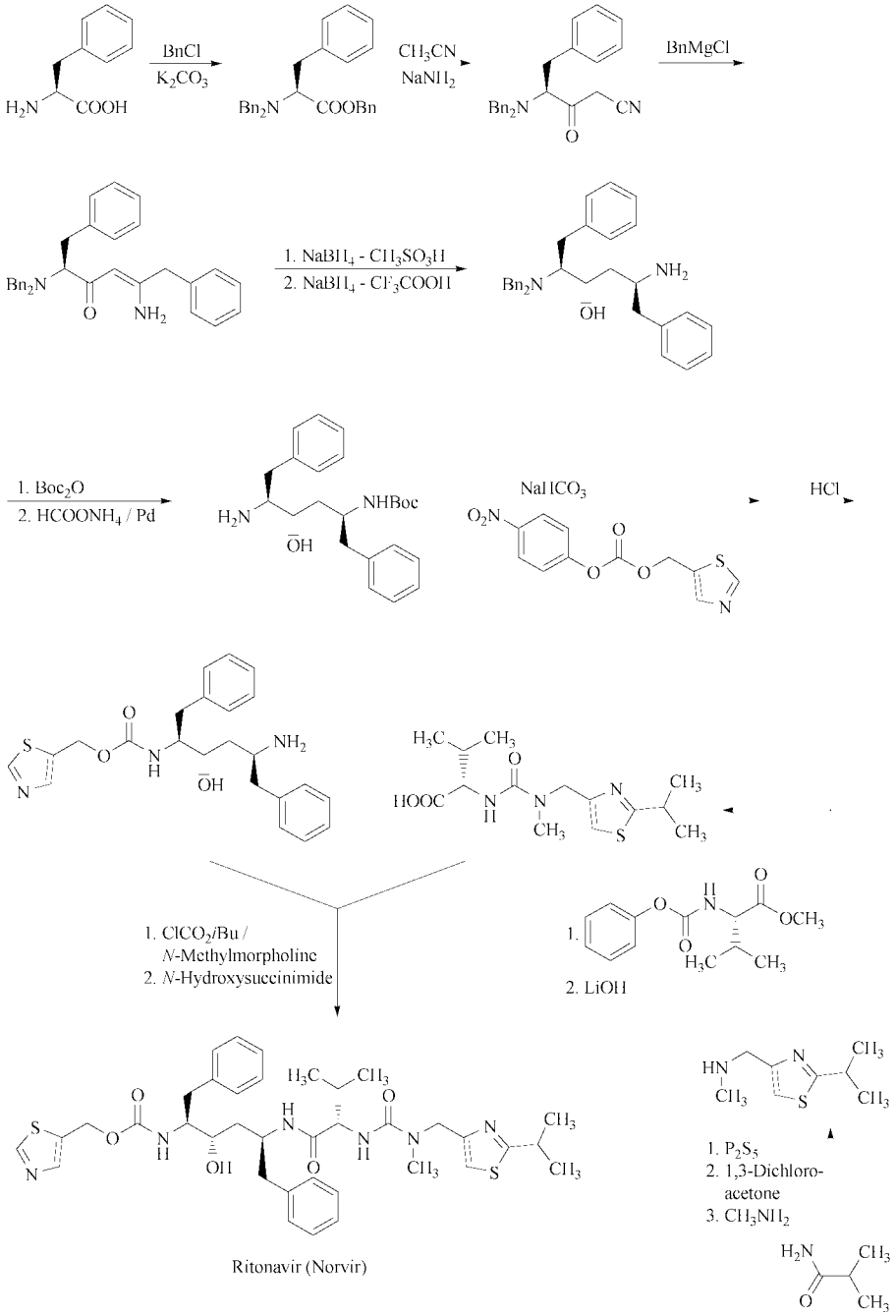

Trade name: Norvir (Abbott Laboratories).

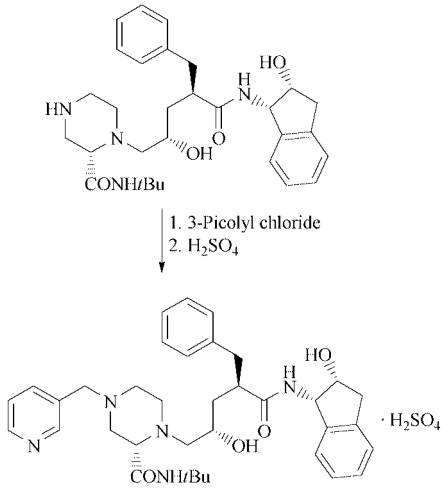

Indinavir sulfate (Crixivan)

Trade name (indinavir sulfate): Crixivan® (Merck & Co. USA).

Lopinavir [*192725-17-0*], [1*S*-[1*R**,(*R**), 3*R**, 4*R**]]-N-[4[[(2,6-dimethylphenoxy)acetyl]amino]-3-hydroxy-5-phenyl-1-(phenylmethyl)pentyl]tetrahydro-alpha-(1-methylethyl)-2-oxo-1(2H)-pyrimidineacetamide; $C_{37}H_{48}N_4O_5$, M_r 628.80, crystals from ethyl acetate/heptane *mp* 124–127 °C;), $[\alpha]_D^{22}$ −22.85° (*c* = 0.4, CH_3OH). Lopinavir is a potent inhibitor of HIV-1 protease whose pharmacokinetic parameters were found to be greatly enhanced by low dose ritonavir. It is indicated for HIV therapy only as a combination with ritonavir.

Synthesis [74] (see right column):

Trade name (lopinavir/ritonavir): Kaletra® (Abbott Laboratories, a 4:1 coformulation tablet or oral solution of lopinavir and ritonavir as a boosting agent).

Atazanavir [*198904-31-3*], (3*S*,8*S*,9*S*,12*S*)-3,12-bis(1,1-dimethylethyl)-8-hydroxy-4,11-dioxo-9-(phenylmethyl)-6-[[4-(2-pyridinyl)phenyl]methyl]-2,5,6,10,13-pentaazatetradecanedioic acid dimethyl ester; free base $C_{38}H_{52}N_6O_7$, M_r 704.9; bisulfate [*229975-97-7*], $C_{38}H_{52}N_6O_7 \cdot H_2SO_4$, M_r 802.9, $[\alpha]_D^{22}$ −46.1 (*c* = 1, 1:1 MeOH/H_2O, pH = 2.6) [75], crystals from acetone *mp* 198–199 °C (decomp.) [76]. Atazanavir is an azapeptide inhibitor of HIV protease. Atazanavir bisulfate is co-administered with low-dose ritonavir in treatment-experienced patients and without ritonavir in treatment-naive patients.

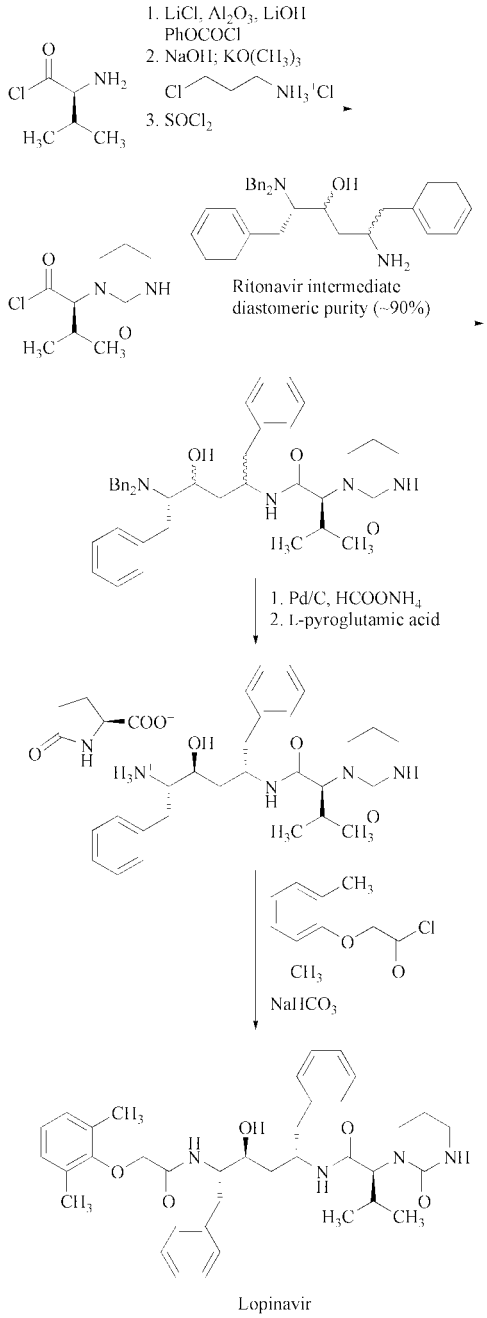

Lopinavir

Synthesis [75]:

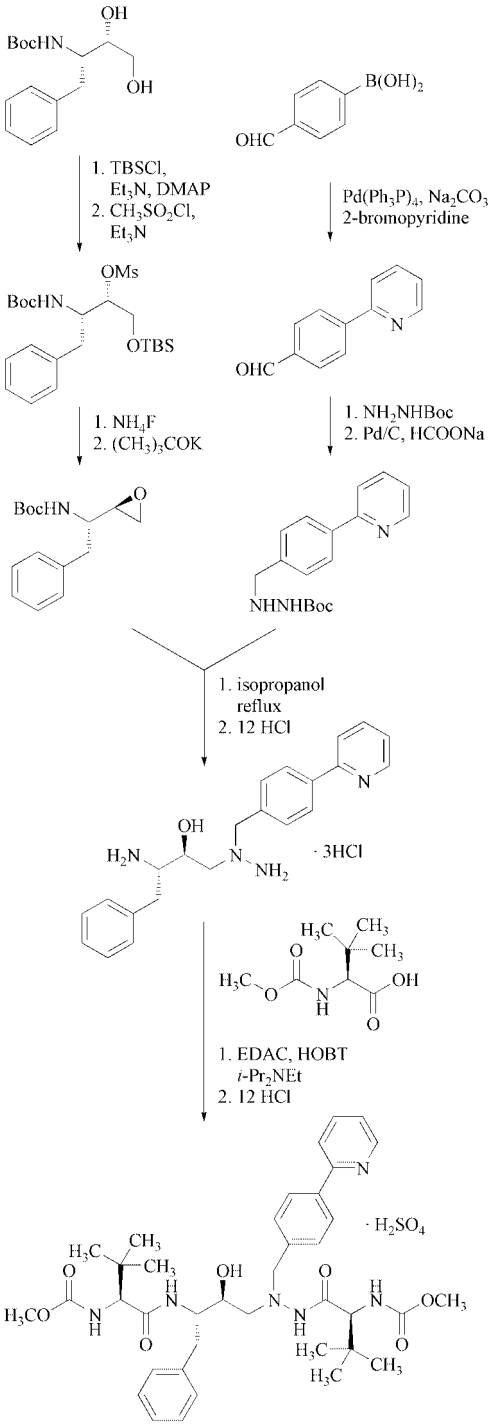

Atazanavir bisulfate (Reyataz)

Trade name (atazanavir sulfate): Reyataz® (Bristol-Myers Squibb Co, licensed from Novartis)

3.4. Non-Peptidic Inhibitors

Tipranavir [*174484-41-4*], *N*-(3-{(1*R*)-1-[(6*R*)- 4-hydroxy-2-oxo-6-(2-phenylethyl)-6-propyl-5,6-dihydro-2*H*-pyran-3-yl]propyl}phenyl)-5-(trifluoromethyl)pyridine-2-sulfonamide; $C_{31}H_{33}F_3N_2O_5S$, M_r 602.7; crystals from ethyl acetate / heptane *mp* 86-89 °C [77] $[\alpha]_D^{22}$ +22 ° (c = 0.96, CH_3OH) [78]. Tipranavir has been approved in 2005 and is one of the latest protease inhibitor to enter the market. It must be administered with low-dose ritonavir.

Synthesis [77]: see next page.

3.5. Compounds in Development

Several second-generation HIV protease inhibitors are presently being evaluated in the clinic (Fig. 3). Among the PIs in advanced stages of development, one can list: Darunavir (TMC114) [*206361-99-1*] by Tibotec-Virco/Johnson& Johnson (approved in 2006 as Prezista™) [79], Brecanavir (GW640385, GW0385, VX-385 another hydroxyethylamine nonpeptidic inhibitor) [*313682-08-5*] by GlaxoSmithKline [80], AG-001859 (an hydroxyethylene isostere containing the unnatural amino acid allophenylnorstatine) [*478410-84-3*] by Pfizer [81, 82].

Trade name (tipranavir): Aptivus® (Boehringer Ingelheim, licensed from Pharmacia & Upjohn Co., must be co-administered with low dose ritonavir as a boosting agent).

Synthesis [77] :

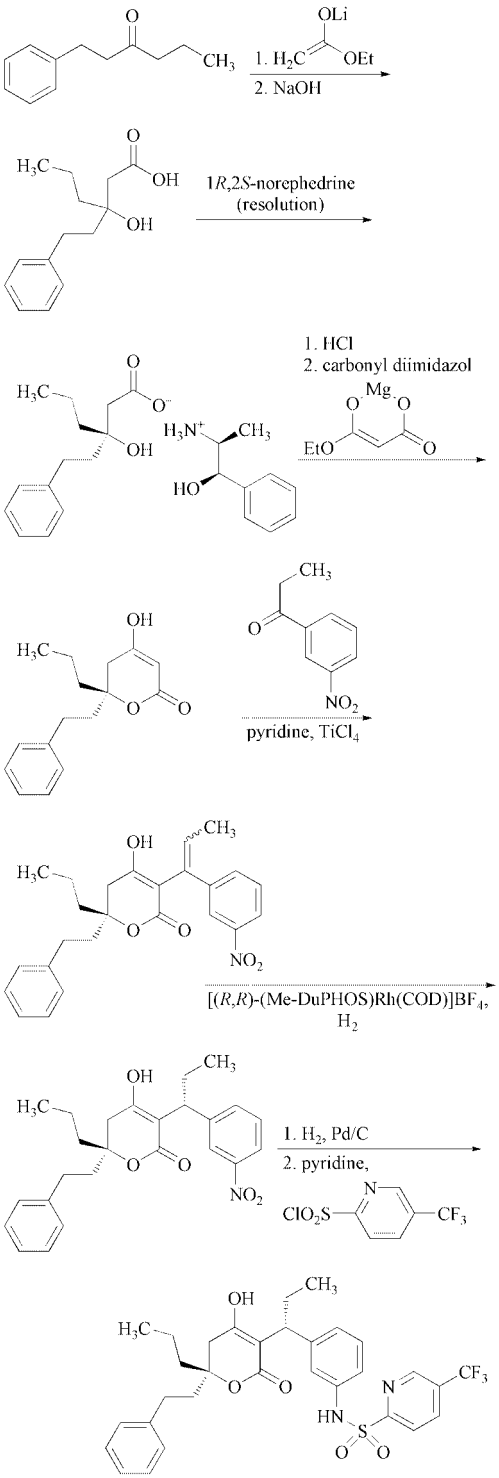

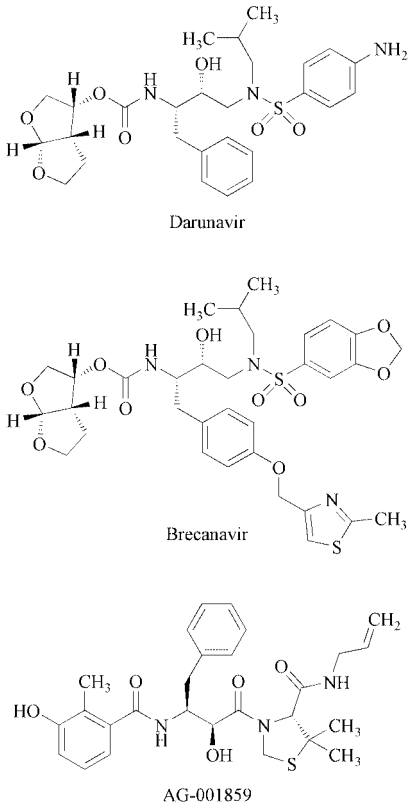

Figure 3. HIV protease inhibitors in clinical development

4. Entry Inhibitors

4.1. Introduction

The process of entry of the HIV virus into the target cell is a complex, multistep event. This process is often divided into two main steps; *attachment* (whereby viral gp120 interacts with CD4 receptor and subsequently with co-receptors CCR5 or CXCR4), and *fusion* (whereby trimeric viral gp41 undergoes conformational changes to initiate membrane fusion). Intensive studies are conducted to identify inhibitors at each of these steps [83, 84]. So far, only one inhibitor of the fusion step has entered the market.

4.2. Fusion Inbihitor

The folding onto each other of the N- and C-peptide regions of gp41 trimer leads to a six-helix bundle which is at the center of the conformational change initiating membrane fusion. Synthetic peptides derived from the C-peptide region have been shown to be potent inhibitors of HIV-1 replication by interfering with the formation of the fusogenic six-helix bundle. They provided the impetus for the discovery of the only fusion inhibitor currently approved so far [85, 86].

Enfuvirtide [*159519-65-0*], N-acetyl-L-tyrosyl-L-threonyl-L-seryl-L-leucyl-L-isoleucyl-L-histidyl-L-seryl-L-leucyl-L-isoleucyl-L-α-glutamyl-L-α-glutamyl-L-seryl-L-glutaminyl-L-asparaginyl-L-glutaminyl-L-glutaminyl-L-α-glutamyl-L-lysyl-L-asparaginyl-L-α-glutamyl-L-glutaminyl-L-α-glutamyl-L-leucyl-L-leucyl-L-α-glutamyl-L-leucyl-L-α-aspartyl-L-lysyl-L-tryptophyl-L-alanyl-L-seryl-L-leucyl-L-tryptophyl-L-asparaginyl-L-tryptophyl-L-phenylalaninamide, T-20; $C_{204}H_{301}N_{51}O_{64}$, M_r 4491.56.

Enfuvirtide was the first fusion inhibitor approved for the treatment of HIV infection in 2003. It is a 36-mer oligopeptide that must be reconstituted in water and self-administered subcutaneously by the patient.

Synthesis [87, 88]: see next page.

4.3. Compounds in Development

Several small-molecule inhibitors, targeting the different steps of the entry process, are currently in advanced development. Among this growing class, one can list:

BMS-488043 (a small molecule blocking viral entry by binding to the viral gp120 and preventing interaction with the CD4 receptor) [*452296-83-2*] by Bristol-Myers Squibb Co [89], Maraviroc (UK-427,857, a CCR5 co-receptor antagonist) [*376348-65-1*] by Pfizer [90], Aplaviroc (GW873140, ONO-4128, AK-602, a CCR5 co-receptor antagonist) [*461443-59-4*] by ONO Pharmaceuticals licensed to GSK [91], Vicriviroc (SCH-D, SCH-417690, a CCR5 co-receptor antagonist) [*306296-47-9*] by Schering-Plough [92] and BlockAide/CR (a peptide-based drug that inhibits the interaction of the V3 loop with host cell lipids) by ADVENTRX Pharm./Synteract Inc. Aplaviroc has recently been discountinued.

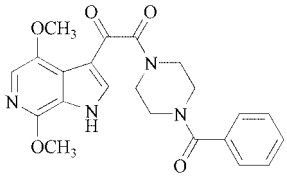

BMS-488043

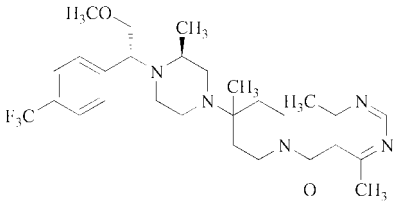

Vicriviroc

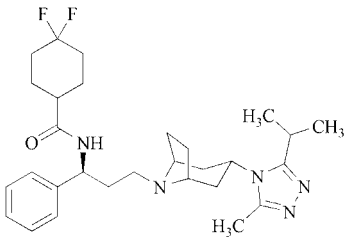

Maraviroc

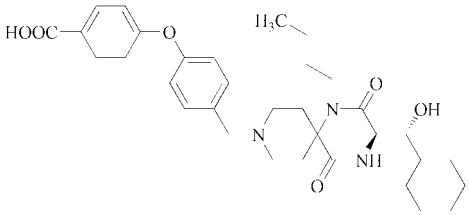

Aplaviroc

5. Emerging Concepts in AIDS Therapy

Despite evidence that the immune response cannot destroy HIV, it is believed that boosting of the body's defense mechanism in preparation for the viral attack, may succeed in containing the virus. Important efforts are being directed at the development of vaccines that can protect against HIV

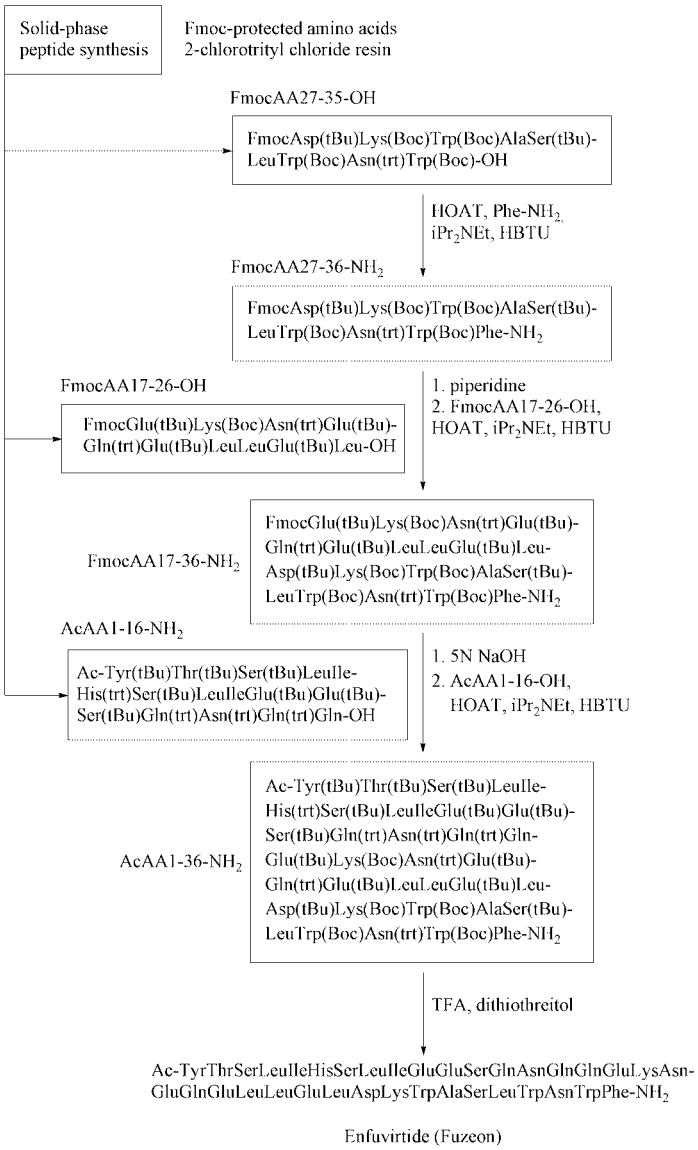

Trade name (enfuvirtide): Fuzeon™ (Roche Laboratories Inc., licensed from Trimeris Inc.).

infection [93–95]. Current strategies include the search for agents that stimulate humoral cellular immune responses by targeting the *env* surface glycoprotein of HIV. On another front, the use of live-attenuated HIV to elicit an immune response is being hampered by potential reversion to pathogenic strains. The lack of understanding of the immune correlates of protection makes vaccine development highly complex and challenging. Coupled to the high variability of the HIV virus, it is unlikely that suitable vaccines for use in humans will become available in the near future [96–98].

Most efforts to treat HIV infection are concentrated on the development of drugs, which arrest the replication cycle of the virus. Most utilized at the present time, reverse transcriptase and protease inhibitors which have long been the major focus of the search for antivirals. Due to the emergence of drug resis-

tance [99], drug failure, and side effects associated with the use of such agents, continued efforts are being directed at the discovery of novel therapies, which target other essential features of the virus' life cycle. These include (i) the viral enzyme integrase which has long been a focus of attention and is now generating promising inhibitors [100–103], (ii) the viral maturation process [104], and more recently (iii) the viral infectivity factor (Vif) [105, 106]. In addition, immune-based approaches such as monoclonal antibodies and immune reconstitution [107–110], inhibitors of cytokine-mediated virus activation [111, 112], soluble CD4 [113] and antisense oligonucleotides [114–116] are all being actively explored.

While major advances in treatment based on combination therapies with reverse transcriptase and protease inhibitors [16] have been made, a definitive cure to the disease remains elusive at this time. A most sought vaccine combined with easier access to drugs in the developing world and strong preventive measures remain the only true hope for arresting this global epidemic.

6. Acknowledgement

We are grateful to Helene Gagnon and Georges Medawar for help with literature searches.

7. References

1. F. Barré-Sinoussi et al., *Science* **220** (1983) 868.
2. R. C. Gallo et al., *Science* **224** (1984) 500.
3. World Health Report 2004, The World Health Organization, http://www.who.int/en/
4. AIDS Epidemic Update 2004, UNAIDS report December 2004, http://www.unaids.org/wad2004/
5. K. M. De Cock et al., *J. Am. Med. Assoc.* **270** (1993) 2083.
6. P. J. Kanki, K. M. De Cock, *AIDS* **8** (Suppl 1) (1994) 85.
7. S. Spira et al., *J. Antimicrob. Chemother.* **51** (2003) 229.
8. A. Tanuri et al., *Antimicrob. Agents Chemother.* **43** (1999) 253.
9. M. M. Thomson et al., *AIDS* **14** (2000) 897.
10. C. A. Andrews, R. A. Koup, *J. Antimicrob. Chemother.* **37** Suppl. B (1996) 13.
11. A. S. Fauci, *Ann. Int. Med.* **114** (1991) 678.
12. A. S. Fauci et al., *Ann. Intern. Med*, **124** (1996) 654.
13. M. Stevenson, M. Bukrinsky, S. Haggerty, *AIDS Res. Hum. Retrov.* **8** (1992) 107.
14. L. Ratner, *Persp. Drug Disc. Des.* **1** (1993) 3.
15. A. Frankel, J. A. T. Young, *Annu. Rev. Biochem.* **67** (1998) 1.
16. E. De Clercq, *Exp. Op. Emer. Drugs* **10** (2005) 241.
17. H. Mitsuya, S. Broder, *Nature* **325** (1987) 773.
18. E. J. Arts, S. F. Le Grice, *Progr. Nucleic Acid Res. Mol. Biol.* **58** (1998) 339.
19. P. L. Darke et al., *Biochem. Bioph. Res. Commun.* **156** (1988) 297.
20. P. A. Darke, J. R. Huff, *Adv. Pharmacol.* **25** (1994) 399.
21. N. E. Kohl et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 4686.
22. R. A. Kramer et al., *Science* **231** (1986) 1580.
23. P. L. Beaulieu et al., *J. Med. Chem.* **40** (1997) 2164.
24. S. Ren, E. J. Lien, *Prog. Drug. Res.* **51** (1998) 1.
25. C. Debouck, *AIDS Res. Hum. Retroviruses* **8** (1992) 153.
26. B. Dunn, *Encyclopedia of Biological Chemistry* **2** (2004) 384.
27. L. Wang, Z. Suo in M. Vaghefi (ed.): *Nucleoside Triphosphates and their Analogs: Chemistry, Biotechnology, and Biological Applications*, CRC Press, London, 2005, p. 207.
28. S. Czernecki, J.-M. Valéry, *Synthesis* (1991) 239.
29. Burroughs Wellcome Co., US 4 724 232, 1988. (J.L. Rideout et al.)
30. V. Bath, E. Stocker, B. G. Ugarkar, *Synth. Commun.* **22** (1992) 1481.
31. J. P. Horwitz, J. Chua, M. Noel, J. T. Donatti, *J. Org. Chem.* **32** (1967) 817.
32. Bristol Myers Squibb Company, US 5 539 099, 1996 (P.M. Skonezny et al.)
33. H. Jin et al., *J. Org. Chem.* **60** (1995) 2621.
34. BioChem Pharma Inc., US 5 696 254, 1997 (T. Mansour et al.)
35. Glaxo Wellcome, WO 98/52949, 1998 (A.C. Brodie et al.)
36. Glaxo Wellcome, US 5 641 889, 1997 (S.M. Daluge, D.A. Livingston)
37. Gilead Sciences, Inc., US 6,723,728, 2004 (Y. Hu, D. Law, K.R. Phares)
38. Triangle Pharmaceuticals, Inc., Emory University, WO 00/09494, 2000 (G.R. Painters et al.)

39. Emory University, US 5,204,466, 1993 (D.C. Liotta, W.-B. Choi)
40. Gilead Sciences, Inc., US 5,935,946, 1999 (J.D. Munger, J.C. Rohloff, L.M. Schultze)
41. K. D. Hargrave et al., *J. Med. Chem.* **34** (1991) 2231.
42. D. L. Romero, *Drugs Future* **19** (1994) 238.
43. Upjohn Company, US 5 563 142, 1996 (J.R. Palmer et al.)
44. M. E. Pierce et al., *J. Org. Chem.* **63** (1998) 8536.
45. L. A. Sorbera, J. Castañer, M. Bayés, *Drugs Future* **27** (2002) 1131.
46. L. J. Stuyver et al., *Antiviral Ther.* **9** (2004) 529.
47. C. Herzmann et al., *Antimicrob. Agents Chemother.* **49** (2005) 2828.
48. M. Zhu, *Curr. Opin. Invest. Drugs* **4** (2003) 1004.
49. S. L. Davies et al., *Drugs Future* **30** (2005) 462.
50. P. A. Janssen et al., *J. Med. Chem.* **48** (2005) 1901.
51. L. A. Sorbera, J. Castañer, M. Bayés, *Drugs Future* **28** (2003) 1149.
52. SmithKline Beecham Corporation, WO 02/070470, 2002 (J.H. Chan)
53. Boehringer Ingelheim, US 6,420,359, 2002 (B. Simoneau)
54. Boehringer Ingelheim, WO 04/054586, 2004 (M.G. Cordingley)
55. A. Wlodawer, J. W. Erickson, *Annu. Rev. Biochem.* **62** (1993) 543.
56. M. Boffito et al., *Curr. Opin. Infect. Dis.* **18** (2005) 1.
57. J. D. Scott, D. James, *Am. J. Health-Syst. Pharm.* **62** (2005) 809.
58. N. A. Roberts, *Int. Antiviral News* **3** (1995) 2.
59. G. Moyle, *Expert. Opin. Invest. Drugs* **5** (1996) 155.
60. S. L. Nightingale, *J. Am. Med. Assoc.* **275** (1996) 273.
61. W. Göhring et al., *Chimia* **50** (1996) 532.
62. T. Inaba et al., *J. Org. Chem.* **63** (1998) 7582.
63. Agouron Pharmaceuticals Inc., WO 98/09951, 1998 (K.R. Whitten, M.E. Deason)
64. Agouron Pharmaceuticals Inc., 5 484 926, 1996, (B. A. Dressmann et al.).
65. Hoffmann-La Roche Inc., US 5 256 783, 1993, (S. Gokhale, M. Schlageter).
66. Vertex Pharmaceuticals Inc., US 5 783 701, 1998 (R.D. Tung, M.A. Murcko, G.R. Bhisetti)
67. Glaxo Group Limited, WO 00/04033, 2000 (I.G. Armitage, A.D. Searle, H. Singh)
68. Glaxo Group Limited, WO 01/00635, 2001 (A.D. Searle)
69. D. J. Kempf et al., *J. Med. Chem.* **41** (1998) 602.
70. A. R. Haight et al., *Org. Process Res. Dev.* **3** (1999) 94.
71. Abbott Laboratories, US 5 567 823, 1996 (J.-H.J. Tien, J.A. Menzia, A.J. Cooper)
72. Abbott Laboratories, WO 94/14436, 1994 (D.J. Kempf et al.)
73. P. J. Reider, *Chimia* **51** (1997) 306.
74. E. J. Stoner et al., *Org. Process Res. Dev.* **4** (2000) 264.
75. Z. Xu et al., *Org. Process Res. Dev.* **6** (2002) 323.
76. Bristol-Myers Squibb, US 6,087,383 (J. Singh, M. Pudipeddi, M.D. Lindrud)
77. Pharmacia & Upjohn Company, WO 00/55150 (B.D. Hewitt, M.J. Burk, N.B. Johnson)
78. T. M. Judge et al., *J. Am. Chem. Soc.* **119** (1997) 3627.
79. L. A. Sorbera, J. M. Bayés, *Drugs Future* **30** (2005) 441.
80. M. H. Hanlon et al., *Biochemistry* **43** (2004) 14500.
81. J. Hammond et al., *Antiviral Ther.* **9** (2004) S17.
82. Agouron Pharmaceuticals Inc., US 2005/0171038, 2005 (J.L. Hammond, A.K. Patick)
83. A. Castagna et al., *Drugs* **65** (2005) 879.
84. N. Cammack, *Curr. Opin. Infect. Dis.* **14** (2001) 13.
85. M. Greenberg et al., *Rev. Med. Virol.* **14** (2004) 321.
86. A. Lazzarin, *Exp. Opin. Pharmacother.* **6** (2005) 453.
87. B. L. Bray *Nat. Rev. Drug Discovery* **2** (2003) 587.
88. Trimeris, US 6,015,881, 2000 (M.-C. Kang et al.)
89. Bristol-Mayer Squib, US 2005/215543, 2005
90. J. Barretina et al., *Drugs Future* **30** (2005) 469.
91. J. A. McIntyre, J. Castañer, *Drugs Future* **29** (2004) 677.
92. J. R. Tagat et al., *J. Med. Chem.* **47** (2004) 2405.
93. A. Ahmad et al., *Curr. HIV Res.* **3** (2005) 261.
94. D. E. Mosier, *D. E. Curr. Drug Targets: Infect. Disord.* **5** (2005) 171.
95. P. L. Nara, G. Lin, *Curr. Drug Targets: Infect. Disord.* **5** (2005) 157.
96. M. Singh, *Virol. J.* **3**:60 (2006).
97. D. R. Burton, J. P. Moore, *Nature Medicine*, vaccine supplement **4** (1998) 495.

98. C. Heilman, D. Baltimore, *Nature Medicine*, vaccine supplement **4** (1998) 532.
99. G. J. Moyle, *Immun. Infect. Dis.* **5** (1995) 170.
100. M. Witvrouw et al., *Curr. Med. Chem: Anti-Infect. Agents* **4** (2005) 153.
101. S. B. Singh et al., *Drugs Future* **30** (2005) 277.
102. Y. Pommier, A. A. Johnson, C. Marchand, *Nat. Rev. Drug Discovery* **4** (2005), 236.
103. D. J. Hazuda et al., *Proc. Natl. Acad. Sci. USA* **101** (2004) 11233.
104. F. Li et al., *Proc. Natl. Acad. Sci. USA* **100** (2003) 13555.
105. K. L. Jones, J. Mak, *Curr. Pharmacogenomics* **3** (2005) 97.
106. K. Stopak, W. C. Greene, *Curr. Opin. Invest. Drugs* **6** (2005) 141.
107. A. Bandera, G. Marchetti, A. Gori, *Exp. Opin. Ther. Patents* **15** (2005) 1115.
108. D. R. Kuritzkes et al., *Journal of Infectious Diseases* **189** (2004) 286.
109. G. Poli et al. in M. Kotb, T. Calandra (eds.): *Cytokines and Chemokines in Infectious Diseases Handbook*, Humana Press, Totowa, 2003, pp. 345–359.
110. A. Trkola et al., *J. Virol.* **75** (2001) 579.
111. N. Cammack, *Antiviral Chem. Chemother.* **10** (1999) 53.
112. S. W. Hunt, III, G. J. LaRosa, *Ann. Rep. Med. Chem.* **33** (1998) 263.
113. J. M. Jacobson et al., *Antimicrob. Agents Chemother.* **48** (2004) 423.
114. R. F. Rando, *Curr. Res. Mol. Ther.* **1** (1998) 67.
115. D. S. Reddy, *Drugs Today* **32** (1996) 113.
116. N. Jing, X. Xu, *Curr. Drug Targets: Infect. Disord.* **1** (2001) 79.

Hydrocyclones

Enrique Ortega-Rivas, Food and Chemical Engineering Program, Autonomous University of Chihuahua; University Campus I, Chihuahua, Chih., 31170, Mexico

1. Introduction 3
1.1. General Remarks 3
1.2. Principle of Operation 3
2. Categories and Application 4
2.1. Thickening 4
2.2. Clarification 4
2.3. Classification 4
2.4. Other Applications 4
3. Theoretical Background 4
3.1. Simultaneous Flow of Fluids and Solids 4
3.1.1. Dynamics of Particles Submerged in Fluids 5
3.1.2. Rheology of Suspensions 6
3.1.3. Flow of Suspensions 7
3.2. Flow Patterns in Hydrocyclones . . 8
3.3. Mechanisms of Particle Separation 9
3.3.1. Equilibrium-Orbit Theory 10
3.3.2. Residence-Time Theory 10
3.3.3. Crowding Theory 10
3.3.4. Turbulent Two-phase Flow Theory . . 10
3.4. Characteristics of Performance of Hydrocyclones 11
3.4.1. Separation Efficiency, Cut Size 11
3.4.2. Capacity, Pressure Drop 16
4. Effect of Variables on Hydrocyclone Performance 16
4.1. Operation Variables 16
4.2. Design Variables 17
5. Selection and Design of Hydrocyclone Systems 17
5.1. Analytical Solutions 18
5.2. Graphical Solutions 18
5.3. Manufactures' Choice 18
5.4. Dimensionless Scale-Up of Hydrocyclones 19
5.4.1. Introduction 19
5.4.2. Definition and Derivation of Dimensionless Groups for Hydrocyclones 19
5.4.3. Scale-Up at Low Concentrations . . . 20
5.4.4. Scale-Up at High Concentrations . . . 21
5.4.5. Considerations for non-Newtonian Behavior 21
5.4.6. Empirical Models 22
5.4.7. Practical Applications 23
6. References 26

Notation

(Dimensions given in terms of mass, M, length, L, time, T, and temperature, θ)

A	arithmetic mean, acceleration (L/T^2)
A	area (L_2)
A_i	area of inlet nozzle of hydrocyclone (L_2)
A_{1n}	parameter dependent on the flow behavior index
C	volume fraction of solids in suspension
C_D	drag coefficient
C'_n	parameter dependant on the flow behavior index
C_{y50}	dimensionless cyclone number
D	pipe diameter (L)
D_c	hydrocyclone diameter (L)
D_i	inlet diameter of hydrocyclone (L)
D_o	overflow pipe diameter of hydrocyclone (L)
D_u	underflow pipe diameter of hydrocyclone (L)
E_p	partial efficiency
E_t	total efficiency
E'_t	reduced total efficiency
Eu	Euler number
f	fanning friction factor
f_{pv}	purely viscous fanning friction factor
$f(x)$	frequency
$F_c(x)$	cumulative percentage oversize of separated solids
F_D	drag force (MLT^{-2})
$F_f(x)$	cumulative percentage oversize of non-separated solids
$F(x)$	cumulative percentage oversize of feed solids

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.c13_c02.pub2

g	acceleration due to gravity (L/T^2)
$G(x)$	grade efficiency
$G'(x)$	reduced grade efficiency
H	height of liquid column or liquid head (L)
H_i	height of rectangular inlet channel of hydrocyclone (L)
$H_{25/75}$	sharpness index
k_p	empirical constant for a family of geometrically similar hydrocyclones
k_1, k_2	constants
K	constant, correlation constant of the power law
K'	fluid consistency index (MT^n/L^2)
l	vortex finder length (L)
L	hydrocyclone length (L)
L_1	length of cylindrical part of hydrocyclone (L)
m	exponent, mass (M)
M	mass flows rate of solids in suspension (M/T)
M_c	mass flows rate of separated solids (M/T)
M_f	mass flows rate of nonseparated solids (M/T)
n	exponent; slope of the curve representing the power law
n_p	empirical constant for a family of geometrically similar hydrocyclones
n'	flow behavior index
N	revolutions per minute or per second (1/T)
O	volumetric flow rate of overflow suspension (L^3/T)
P	pressure (M/LT^2)
q	volumetric flow rate per unit area (L/T)
Q	volumetric flow rate (L^3/T)
r	parameter function of a log-ln plot in the coarse region
R	radius of rotation (L)
R_{av}	average radius of rotation (L)
Re	Reynolds number
Re_p	particle Reynolds number
Re^*	Reynolds number for power-law fluids
R_f	underflow-to-throughput ratio
s	parameter function of the slope of the log-ln plot in fine size regions
Stk_{50}	Stokes number
Stk'_{50}	Stokes number including the reduced cut size
Stk^*_{50}	Stokes number for power-law fluids
$Stk^*_{50}(r)$	Stokes number for power-law fluids including the reduced cut size
t	time (T)
t_d	detention time (T)
t_s	settling time (T)
T	temperature (θ)
u	fluid-particle relative velocity (L/T)
u_t	terminal settling velocity under gravity (L/T)
U	volumetric flow rate of underflow suspension (L^3/T)
v	linear velocity, hydrocyclone characteristic velocity (L/T)
v_g	terminal settling velocity under gravity (L/T)
v_i	inlet tangential velocity (L/T)
v_o	maximum tangential velocity (L/T)
v_r	radial settling velocity (L/T)
v_t	tangential velocity (L/T)
V	volume (L^3)
w	vertical velocity at cyclone wall (1/T)
x	particle size (L)
x_a	arithmetic mean of particle size (L)
x_c	cut size (L)
x_g	geometric mean of particle distribution, mass median size of particle (L)
x_m	mode of particle size distribution
x_o	maximum particle size in sample (L)
x_r	constant function of particles size range (L)
x_0	particle size with zero centrifugal efficiency (L)
x_{50}	cut size (L)
x'_{50}	reduced cut size (L)
x_{98}	approximate particle size fully separated (L)
Z	acceleration factor
Z_i	width of rectangular inlet channel of hydrocyclone (L)
	Greek
α	constant
γ	viscosity coefficient for power-law fluids (M/LT^{2-n})
$\dot{\gamma}$	shear rate (1/T)
Δ	difference
θ	hydrocyclone cone angle
μ	liquid absolute viscosity (M/LT)
μ_o	viscosity of pure solvent (M/LT)
ρ	liquid density (M/L^3)
ρ_m	slurry density (M/L^3)
ρ_s	solids density (M/L^3)
σ	standard deviation

τ	shear stress (M/LT^2)
τ_w	shear stress at tube wall (M/LT^2)
ϕ	volume fraction of spheres in suspension
ω	angular velocity (1/T)

1. Introduction

1.1. General Remarks

Hydrocyclones have been less employed than centrifuges to develop high centrifugal forces and perform separation of solids. Although hydrocyclones were first patented more than 100 years ago, industrial use of them was practically nonexisting until the late 1940s. Hydrocyclones were early utilized in the paper and mineral industries, to become later popular in a larger number of factories from chemical, cement, and nuclear to the food, pharmaceutical, and oil industries. Although hydrocyclones continue finding new applications, the dearth of specialized literature about them is surprising. With only two known books written in English since the 1980s and some attention paid to them in the literature, there is more to learn about hydrocyclones than about other solid–liquid separation techniques.

1.2. Principle of Operation

A hydrocyclone is a sedimentation device, which employs enhanced gravity force to separate solid from a carrying liquid. This mentioned force is in fact the centrifugal field produced when a fluid suspension is pumped tangentially into a cone-cylindrical body. As a vortex is created, coarse particles move towards the wall and fines towards the axis. Flow is at first down the inner wall of the cylindrical section and the cone, as far as the stagnation point near the cone apex. There, because the opening is small, the downward primary vortex is forced to turn upwards again, thus forming the secondary vortex. Thus, at the bottom of the cone the flow splits into two streams, the underflow containing most of the coarse particles, and the overflow with most of the fines in it. The overflow pipe generally protrudes into the cylindrical body, to form what is known as the vortex finder. A diagram of a hydrocyclone is shown in Figure 1. As can be concluded, separation is based on difference in settling velocities, as in sedimentation. Thus, size, density and shape of particles play a role important in the operation.

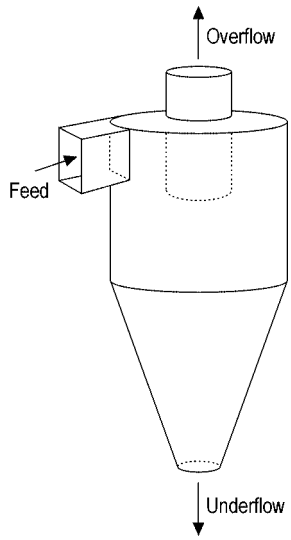

Figure 1. Diagram of a hydrocyclone

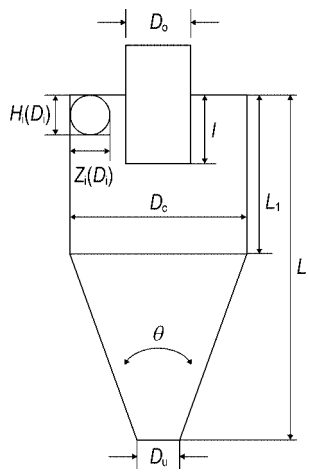

Figure 2. Nomenclature and dimensions of hydrocyclone D_c: cyclone body diameter, D_o: overflow pipe diameter, D_i: inlet diameter, $[(4HiZ_i)/\pi]^{1/2}$ for rectangular inlet channels, l: vortex finder length, L_1: cylindrical section length, L: hydrocyclone length, θ: cone angle, D_u: underflow orifice

Typical sizes of hydrocyclones range from 10 mm to 2.5 m in diameter, capacities of sin-

gle units go from 0.1 to 7200 m^3/h, and pressure drops vary from 0.34 to 6 bar. The efficiency, reported as cut size, lies between 2 and 250 µm. The main advantages of the hydrocyclone are its simple, compact, and inexpensive construction, its ease of operation (no moving parts), its short residence time, and its manageability of sizes finer than those treated in mechanical classifiers. Some disadvantages are its high power consumption, its high wear, its inflexibility and its problem with blockage, especially in small units. Figure 2 presents typical dimensions and standard nomenclature of hydrocyclones.

2. Categories and Application

Hydrocyclones are very versatile pieces of equipment. They can be employed as thickeners, clarifiers, classifiers and even in liquid–liquid separations. A quite useful and up-to-date guide of applications and models of hydrocyclones is given in [1].

2.1. Thickening

For thickening proposes, hydrocyclones have proved to be efficient, with underflow concentration as high as 50 % obtained with some material. Thus, they can easily compete in this field with gravity thickeners, with advantages of less space requirement. The main problem in the operation, however, is the risk of blockage, but hydrocyclones with variable underflow orifice are employed to avoid this difficulty.

2.2. Clarification

When used in clarification, small diameter hydrocyclones with parallel multicyclone arrangement are preferred. Multiple cyclone units with six to several hundred hydrocyclones in parallel are available and have been successfully used.

2.3. Classification

Since the principle of operation of hydrocyclones is centrifugal, they can be used to separate solids from solids in a suspension according to particle size. Sometimes, either coarse or fine particles may be removed from the product, and hydrocyclones have performed such operations known as degritting or refining, and desliming or washing, respectively. This classifying characteristic of hydrocyclones is also employed to improve the performance of other filtration or separation equipments. The coarse and fine particles separated from hydrocyclones normally follow different paths in a processing plant, being sometimes fed into other units, to finish as a different product.

2.4. Other Applications

As hydrocyclones have been used either as thickeners or as clarifiers, they have also performed both functions in one operation. In order to do so, two or more units are connected in series. One of them functions as thickener and the others as clarifiers. A typical arrangement, which operates with the thickener at the first stage, followed by two clarification stages, and the underflow of them returned to the feed, has been widely described in [2]. Other arrangements that have also maximized the overall performance have been analyzed in [1]. The centrifugal features of hydrocyclones give them possibilities to perform separation of two immiscible liquids. Hydrocyclones have been applied in separation of oil from water, in dewatering of light oils and in producing highly concentrated samples of lighter dispersed phases. Other applications such as countercurrent washing of solids and separation of gas bubbles from feed liquids have as well been tried, giving satisfactory results.

3. Theoretical Background

3.1. Simultaneous Flow of Fluids and Solids

As hydrocyclones are usually fed with a dispersion of solids in liquid, a study of the principal characteristics of suspension is necessary. Since concentration can be vary widely in hydrocyclone operations, particle–fluid interaction, as well as proper rheology of the whole suspension,

are of the utmost importance. For a detailed description of fluid dynamics and rheology see → Fluid Mechanics.

3.1.1. Dynamics of Particles Submerged in Fluids

If a particle moves relative to the fluid in which it is suspended, the force opposing the motion is known as the drag force. Knowledge of the magnitude of this force is essential if the particle motion is to be studied. Conventionally, the drag force F_{D} is expressed as:

$$F_{\mathrm{D}} = C_{\mathrm{D}} A \frac{\rho u^2}{2} \tag{1}$$

where u is the particle–fluid relative velocity, ρ is the fluid density, A is the area of the particle projected in direction of the motion, and C_{D} is a coefficient of proportionality known as the drag coefficient. Assuming the drag force is due to the inertia of the fluid, C_{D} would be constant and dimensional analysis shows that C_{D} is generally a function of the particle Reynolds number, i.e.:

$$Re_{\mathrm{p}} = \frac{ux\rho}{\mu} \tag{2}$$

where x is the particle size and μ the medium viscosity; the form of the function depends on the regime of the flow. This relationship for rigid spherical particles is shown in Figure 3. At low Reynolds number under laminar flow conditions when viscous forces prevail, C_{D} can be determined theoretically from Navier–Stokes equations (see → Fluid Mechanics, Section 3.2) and the solution is known as Stokes' law and represented by:

$$F_{\mathrm{D}} = 3\pi\mu ux \tag{3}$$

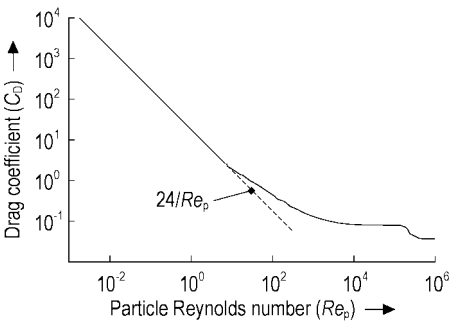

Figure 3. Drag coefficient versus particle Reynolds number for spherical particles

This is an approximation, which gives the best results for $Re_{\mathrm{p}} \to 0$; the upper limit of its validity depends on the error that can be accepted. The usually quoted limit for the Stokes region of $\mathrm{Re_p} = 0.2$ is based on an error of about 2 % in the terminal settling velocity. Equations 1, 2 and 3 combined give another form of Stokes' law as follows:

$$C_{\mathrm{D}} = \frac{24}{Re_{\mathrm{p}}} \; (Re_{\mathrm{p}} < 0.2) \tag{4}$$

For Reynolds numbers > 1000 the flow is fully turbulent with inertial forces prevailing, and C_{D} becomes constant and equal to 0.44 (the Newton region). The region in-between $\mathrm{Re_p} = 0.2 - 1000$ is known as the transition region, and C_{D} is either described in a graph or by one or more empirical equations.

For a particle of mass m under the influence of a field of acceleration a, the equation of motion is:

$$m\left[\frac{\mathrm{d}u}{\mathrm{d}t}\right] = ma - ma\left[\frac{\rho}{\rho_{\mathrm{s}}}\right] - F_{\mathrm{D}} \tag{5}$$

where ρ_s is the density of the solids and t is the time.

In applications of hydrocyclones concerned with separation of fine particles, which are all most difficult to separate, the Reynolds numbers are low, often less than 0.2, due to the low values of x and u. Therefore, it is reasonable also that the time necessary for the particle velocity to reach values very close to the terminal settling velocity is very short and being the field of acceleration gravity, Equation (5) can be solved to give:

$$u_t = \frac{x^2\,(\rho_{\mathrm{s}} - \rho)\,g}{18\mu} \tag{6}$$

in which the acceleration due to gravity g has replaced a and u_{t} is known as the terminal settling velocity under gravity.

The radial settling velocity in a hydrocyclone v_{r} is due to the centrifugal acceleration, which is proportional to the square of the tangential velocity of the particle, and indirectly proportional to the radius of the particle position. As the tangential motion of particle is unopposed, the tangential particle velocity can be taken as equal to the tangential component of the fluid

velocity at the same point. The inclusion of the centrifugal terms in Equation (6) leads simply to:

$$v_{\mathrm{r}} = \frac{x^2\,(\rho_{\mathrm{s}}-\rho)\,R\omega^2}{18\mu} \tag{7}$$

where R is the radius of the particle position and ω is the angular velocity.

As the concentration of the suspension increases, particles get closer together and interfere with each other. If the particles are not disturbed uniformly, the overall effect is a net increase in settling velocity because the return flow caused by volume displacement predominates in particles-sparse regions. This is the well-known effect of cluster formation, which is significant only in nearly monosized suspensions. With most practical widely dispersed suspension, clusters do not survive long enough to affect the settling behavior and, as the return flow is more uniformly distributed, the settling rate steadily declines with increasing concentration. This phenomenon is referred to as hindered settling and can be theoretically approached in three different manners: as a Stokes' law correction by introduction of a multiplying factor; by adopting effective fluid properties for the suspension different from these of the pure fluid; and by determination of bed expansion with a modified version of the well know Carman–Kozeny equation. All the approaches can be shown to yield essentially identical results. Some important correlations accounting for the hindered settling effect have been reviewed in [3]. It was shown that their differences are minimal and, according to this, the Richardson and Zakii equation is an obvious choice in practice. Such relation can be expressed as:

$$\frac{u}{u_{\mathrm{t}}} = (1-C)^{4.65} \tag{8}$$

where u is the settling velocity at concentration C and u_{t} is the settling velocity of a single particle.

The relationship above applies only to free, particulate separation unaffected by coagulation or flocculation, and where all particles are of uniform density.

3.1.2. Rheology of Suspensions

The rheological properties of suspension have been studied since the beginning of the 20th century. The first research was done by Einstein[4] in his classical study of the viscosity of dilute suspensions of rigid spheres. His approach is purely hydrodynamic, and his model consists of an isolated sphere situated in a simple shear flow field in an infinite fluid. A number of workers have been extending Einstein's analysis [3, 5]. Three approaches have been presented: the theoretical basis of viscosity computation, the effect of particle texture and shape, and the effect of concentration. Several relations have been developed for suspensions in with nearest neighbor interactions cannot be neglected. Guth and Simha [6] considered the first order effect of spheres interacting with one another. They found that:

$$\frac{\mu}{\mu_{\mathrm{o}}} = 1+2.5\phi+14.1\phi^2 \tag{9}$$

where μ is the viscosity of the suspension, μ_{o} is the viscosity of the pure solvent and ϕ is the volume fraction of spheres in the suspension. A later correction gave a value of 12.6 for the last constant [7].

Several authors have made experimental studies of the effects of concentration on the viscosity of suspension of spheres. The data they have proposed are considerably scattered. Thomas [8] has tried to find the sources of scatter in the data and found that scattering is caused in large part by variations in particle size. For small particle sizes (diameters $<$ 1–10 µm) colloidal forces become important and the viscosity begins to increase as particle size decreases. The viscosity is also shear rate-dependent in this case. For the particles larger than 1 to 10 µm, the particle Reynolds number becomes significant. The inertial effects results in an increase in relative viscosity with increasing particle size. Thomas has developed a unique curve, by eliminating diameter and shear rate effects. For low concentration data, such a graph can be represented by:

$$\frac{\mu}{\mu_{\mathrm{o}}} = 1+2.5\phi+11.4\phi^2 \tag{10}$$

where all the components are already defined.

The behavior of some specific systems has also been reported in the literature. Lin and Brodkey [9] studied rheological properties of

slurry fuel, concluding that they are shear-thinning under most conditions, but turn shear-thickening at extreme conditions. GRISKEY et al. [10] investigated the rheological behavior of cornstarch dispersions and found that they go from dilatant (shear-thickening) behavior at low shear rates to Newtonian behaviors at higher shear rates. The effect of particle aggregation on suspension has also been studied. DEVRIES [11] presented a number of experimental results concerning the particle aggregation in polymer lattices induced or accelerated by simple shear flow. No stable flow is possible under these conditions, and the apparent viscosity is an increasing function of the time of shearing.

3.1.3. Flow of Suspensions

Some flow patterns of relative interest will be reviewed. One of the most studied cases is that of transport of suspensions such as slurries through pipelines, due mainly to its occurrence in a wide number of industrial processes. Most of the theory developed for flow through tubes applies to laminar flow in smooth tubes. There are some relationships available for pressure-loss flow rate in purely viscous and viscoelastics flows. Relations for thixotropic and rheopectic systems appear not to be available.

Under the assumptions that purely viscous behavior prevails and that no slip occurs at the tube wall, the power-law equation for laminar flow is in the form [12]:

$$\tau_{\mathrm{w}} = \frac{D\Delta P}{4L} = K\prime\left[\frac{8v}{D}\right]^{n\prime} \tag{11}$$

where τ_w is the shearing stress at the wall of the tube, D is the tube diameter, ΔP is the pressure drop, L is the length of the tube, v the mean velocity, n' is the slope of the line when the data are plotted on logarithmic coordinates. For n' = 1, the fluid is Newtonian; for n' <1, pseudoplastic, or Bingham plastic if the curve does not go through the origin; and for n' >1, dilatant. The term K' is the consistency index; as the name suggest, the larger its value the thicker or more viscous the fluid.

The constant K' may be related to the analogous power-law constant K (see → Fluid Mechanics, Section 4.1.2) as follows [13]:

$$K' = K\left[\frac{3n'+1}{4n'}\right]^{n}\left[\frac{8v}{D}\right]^{n-n'} \tag{12}$$

if the fluid obeys the power-law, $n = n$' and:

$$K' = K\left[\frac{3n+1}{4n}\right]^{n} \tag{13}$$

Since Equation (11) rigorously portrays the laminar flow behavior of the fluid (provided n' and K' are evaluated at the correct shear stress), one may use it to define a Reynolds number applicable to all purely viscous fluids under laminar flow conditions. This dimensionless group can be derived simply by the substitution of $(D\Delta P/4L)$ from Equation (11) into the usual definition of the fanning friction factor, i.e.:

$$f = \frac{D\Delta P/4L}{\rho v^2/2} \tag{14}$$

such substitution leads to:

$$f = \frac{16\gamma}{D^{n\prime}v^{2-n\prime}\rho} \tag{15}$$

where $\gamma = K'8^{n'-1}$ and all the remaining component as already defined.

By letting $f = 16/Re$ as for Newtonian fluids in laminar flow, the above mentioned generalized Reynolds number can be obtained as:

$$Re* = \frac{D^{n'}v^{2-n'}\rho}{\gamma} \tag{16}$$

If the equation is desired in terms of K instead of K', Equation (12b) may be substituted into Equation (15) and:

$$Re* = \frac{D^{n}v^{2-n}\rho}{K8^{n-1}\left[\frac{3n+1}{4n}\right]^{n}} \tag{17}$$

For Newtonian fluids $n' = 1$ and $K' = \mu$, so γ reduces to μ and Re^* in Equation (16) transforms to the familiar $Dv/\rho/\mu$ showing that this traditional dimensionless group is merely a special restricted form of the more general described here.

Non-Newtonian fluids in turbulent flow generally show lower friction factors and, consequently, lower pressure drops than Newtonian fluids at corresponding Reynolds number. A detailed analysis of non-Newtonian flow at turbulent regime began with two papers published in 1959 by DODGE and METZNER [14] and SHAVER and MERRIL [15]. DODGE and METZNER used solutions of carboxymethyl cellulose and carbopal (a B. F. Goodrich soluble thickener), as well as clay

suspensions. They presented a viscous correlation in terms of parameters from shear-stress–shear-rate data and concluded that except for the carboxymethyl cellulose, the behavior of these materials could be correlated with the following generalization of the Newtonian friction factor:

$$\sqrt{\frac{1}{f}} = A_{1n}\log\left[\mathrm{Re}*(f)^{1-0.5n\prime}\right] + C_n^* \tag{18}$$

where A_{1n} and C_n' are parameters dependent on the flow-behavior index *n*. While the choice of power-law equation in the derivation requires that the Reynolds number used in Equation (17) be the power-law special case of Equation (16), it has been shown [14] that any errors due to this approximation are less important under turbulent flow conditions than in the laminar region. The utility of Equation (17) and of Reynolds number outside the laminar region has been shown diagrammatically in [14]. The authors found excellent agreement between experimental and extrapolated data. Furthermore, deviations from smooth curves were no greater for those fluids, which did not obey the power law, than for those which did. The parameters A_{1n} and C_n' in Equation (17) must be evaluated empirically, just as in the simpler case of Newtonian behavior.

By far, most of the study of the flow behavior of non-Newtonian suspensions has been devoted to monosized and monoshaped systems, mainly of spherical particle geometry. In reality, however, solid–liquid mixtures usually consist of dissolved solids with a variety of particle sizes, shapes, and concentrations. The available theory for the behavior of non-Newtonian suspensions is, therefore, of a limited application, and when dealing with real systems, empirical corrections are necessary to fully explain the suspension properties.

Studies of some real applications such as slurry transport are also encountered in the literature. The slurry to be transported may have settling or nonsettling characteristics. Sand in water, salt in benzene, and clay in drilling mud are typical examples of settling slurries, while paper pulp is a good example of nonsettling slurry. The law governing the settling rate of particles at low Reynolds numbers is the well know Stokes law. Settling phenomena of slurries in pipelines, as well as modifications of original Stokes law for flocculated suspensions and particles of odd shapes, have been studied by Thomas [16]. The turbulent regime is sometimes preferred to transport slurries within pipe in order to prevent settling, and to provide high throughputs. In this case, the viscosity of the suspension has been normally used to correlate the turbulence friction factors.

The problem of distribution and mixing of solids in turbulent slurry flow has been studied. The magnitude of flow velocity or Reynolds number required for maintaining turbulent mixing of solid–liquid slurries, may be computed from the following equation:

$$\frac{\rho v^2}{(\rho_\mathrm{s}-\rho)\,gx} = K\prime\left[\frac{\rho_\mathrm{m} v D}{\mu}\right]0.775 \tag{19}$$

where ρ is the liquid density, ρ_s the solid density, ρ_m the slurry density, v the minimum standard velocity, g the acceleration due to the gravity, D the mean particle diameter, μ the viscosity and K' a constant. When densities of solids and liquids are known, along with the volume fraction of solids in the slurry, the density of the mixture may be calculated by:

$$\rho_\mathrm{m} = C\rho_\mathrm{s} + (1-C)\,\rho \tag{20}$$

where C is the volume fraction of solids in the slurry. Equation (18) is recommended for mean diameters between 50 and 500 µm.

3.2. Flow Patterns in Hydrocyclones

The flow pattern in a hydrocyclone is complex. However, it is normally accepted that the hydrocyclone fluid flow consists of the three velocity components: tangential, vertical, and radial. Figure 4 shows a diagrammatically representation of these three velocity distributions within a hydrocyclone. With respect to the tangential component v_t, the linear velocity of the feed stream at the inlet point is given by:

$$v_\mathrm{t} = \frac{Q}{A_\mathrm{i}} \tag{21}$$

where Q is the volumetric flow rate and A_i is the area of the inlet nozzle.

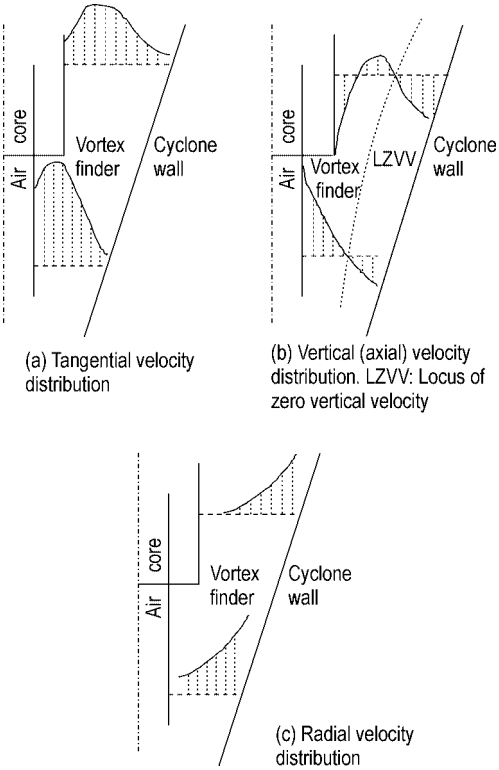

Figure 4. Velocity distribution profiles within a hydrocyclone

As the flow develops in a spiral manner towards the air core, the circular speed increases as the diameter of rotation decreases. Under no friction loss conditions, the angular speed remains constant and is given by:

$$v_{\mathrm{t}} R = \text{constant} \tag{22}$$

where R is any radius of rotation.

In practical systems frictional losses result in a progressive decrease in tangential speed, thus Equation [21] needs to be modified, e.g.:

$$v_{\mathrm{t}} R^{n} = \text{constant} \tag{23}$$

where: $0<n<1$.

Experience has shown that the value of n decreases with increasing solids contents, with a maximum value of about 0.3 for solids contents of about 60 wt %. For normal operating conditions of 15 to 25 % of solids approximately, Trawinski[17] has suggested an average value of 0.5.

The maximum tangential velocity will occur in the secondary vortex immediately adjacent to the air core. Taking the proposed average value of 0.5, this maximum velocity will be theoretically described by:

$$v_{\mathrm{o}} = v_{\mathrm{t}} \left[\frac{D_{\mathrm{i}}}{D_{\mathrm{o}}} \right]^{1/2} \tag{24}$$

Practically, the limit on the maximum tangential velocity is given by pressure drop across the hydrocyclone. Under ideal conditions of no friction loss, the following relationship describes this maximum velocity:

$$v_{\mathrm{o}} = \left[\frac{2\Delta P}{\rho_{\mathrm{m}}} \right]^{1/2} \tag{25}$$

where ΔP is the pressure drop and ρ_{m} is the density of the slurry. As can be seen from Figure 4b, the profile of the vertical velocity component shows a well-defined position of zero velocity outside of which there is a low speed downwards, and inside of which the flow is much higher and upwards. This locus of zero vertical velocity (LZVV) follows the profile of the cyclone.

The radial component of velocity is inward, and accurate measurement of it is difficult because its magnitude is small. Apparently, such speed decreases proportionally to radius decreasing, from a value that can be expressed by:

$$v_{\mathrm{r}} = w \tan \left[\frac{\theta}{2} \right] \tag{26}$$

where w is the vertical velocity at the wall and θ is the angle of the cone.

The radial position at which this velocity falls to zero is not known precisely. Figure 4c shows the radial component of velocity.

3.3. Mechanisms of Particle Separation

According to the flow patterns just described, several models to explain the actual mechanism of particle separation have been developed. Each model works well for particular hydrocyclones only, the models are not of general application, because every one relies on a single dominant feature of the flow inside the hydrocyclone. There have been at least four approaches to describe the behavior of particles in a hydrocyclone: the equilibrium-orbit theory, the residence-time theory, the crowding theory and the turbulent two-phase theory.

3.3.1. Equilibrium-Orbit Theory

An equilibrium orbit is proposed at a radius position, at which the inward drag on the particle due to the inward radial velocity of the fluid counterbalances the outward force to the liquid rotation, i.e.:

$$\frac{\pi x^3}{6}\left(\rho_s - \rho\right)\frac{v_t^2}{R} = 3\pi\mu v_r x \tag{27}$$

or:

$$\frac{x^2}{18}\frac{\left(\rho_s - \rho\right)}{\mu} = \frac{R v_r}{v_t^2} \tag{28}$$

where x is the particle diameter or size, and the densities and velocities are as previously defined. If $v_t \sim (v/R^n)$ and $v_r \sim (v/R^m)$, where v is the inlet velocity, then:

$$\frac{x^2}{18}\left[\frac{\rho_s - \rho}{\mu}\right] = \left[\frac{K}{v}\right] R^{1+2n-m} \tag{29}$$

For given inlet velocity and liquid, the radius of the orbit is proportional to the particle size. If R is greater than the radial position for zero vertical velocity, it is assumed that all particles will be collected. If R is less than the radial position, all particles will be carried away in the underflow. Thus, a particle will have a 50 % chance of removal if the radius of its equilibrium orbit is equal to that of the zero vertical velocity. Using this approach, Bradley [7] derived a relation making assumptions to the position of the locus of zero vertical velocity profile.

3.3.2. Residence-Time Theory

This theory advocates the importance of the time spent in the hydrocyclone. Assuming the spiral has a constant velocity equal to the average inlet velocity, and supposing the liquid follows N complete spirals at an average radius R_{av}, the residence time will be:

$$t = \left[\frac{2\pi R_{av}}{v}\right] N \tag{30}$$

a particle of size x will travel a distance L in this time t where:

$$\frac{L}{t} = v = \frac{x^2\left(\rho_s - \rho\right)}{18\mu}\left[\frac{v^2}{R}\right] \tag{31}$$

Combining Equations 30 and 31 the following relationship is obtained:

$$x^2 = \frac{9\mu L}{N\pi v\left(\rho_s - \rho\right)} \tag{32}$$

This is a form similar to Equation 29 without the dependence on radial positions.

Using direct analogy with gravity settling, Trawinski [22] proposed another simple approach, which may be considered as a variation of the residence-time theory. He reported the following equation:

$$x_c = \left[\frac{18\mu q}{g\left(\rho_s - \rho\right) Z}\right]^{1/2} \tag{33}$$

where x_c is known as the cut point, q equals (Q/A) and Z is the acceleration factor. Including further considerations, e.g., for capacity and the separating area (surface of the cylinder formed by the air core), Trawinski [17] also proposed:

$$x_c = k_1\left[\frac{18\mu}{\left(\rho_s - \rho\right) g}\right]^{1/2}\left[\frac{D_c D_i}{l}\right]^{1/2}(H)^{-0.25} \tag{34}$$

where k_1 is a constant, H is the liquid head, and all the geometric characteristics are as defined in Figure 2.

3.3.3. Crowding Theory

Due originally to Fahlstrom [18], this theory proposes that the cut size is a function of the capacity of the underflow orifice and the particle size distribution. The crowding effect can swamp the primary interaction to the extent that the cut size can be estimated from the mass recovery to the underflow. The support to this theory has been given mainly by Bloor And Ingham [19], by means of the mathematical models developed to describe the fluid mechanics of hydrocyclones.

3.3.4. Turbulent Two-phase Flow Theory

This theory states the importance of the effect caused by turbulent flow in hydrocyclones. According to this, it is proposed that separation takes place under the influence of the centrifugal field and the turbulent transport in the two-phase flow. The influence of turbulence can be described by an eddy diffusion coefficient equal to, or smaller than the diffusion of the fluid particles. The assessment of the influence of such diffusion on separation is very difficult in practice.

Bloor and Ingham [19] have also contributed to the theoretical background of this theory. They derived an analytical form of the radial velocity and found an analytical solution to the equation of motion for the tangential speed. They also developed a more elaborated and physically realistic model, for the eddy viscosity based on the Prandtl mixing theory, which greatly simplified the equation of motion.

Schubert and Neesse [20] proposed a set of equations and derived two models for turbulent cross-flow wet classification: the suspension-partition model and the suspension-tapping model. In the former the flow is divided into overflow and underflow without changes in the total cross-section, whereas in the latter overflow and underflow are tapped from the main flow through small outlet openings. According to them, the suspension-tapping model is analogously applicable to the separation in hydrocyclones.

3.4. Characteristics of Performance of Hydrocyclones

The most important features of performance of hydrocyclones are the particle cut size, the pressure drop and the volume throughput. Previous work has established the cut size as the best way to predict efficiency in hydrocyclones. In order to understand and utilize the cut size concept, it is necessary to deal with at least two topics, i.e., particle size analysis and efficiency of solid–liquid separation (see → Particle Size Analysis and Characterization of a Classification Process; → Solid–Liquid Separation, Introduction).

3.4.1. Separation Efficiency, Cut Size

The solid phase of suspensions to be treated in solid–liquid separation equipment generally consists of an immense number of particles of diverse sizes and shapes. All this population of particles needs to be identified or characterized. The frequency of occurrence of particles of every size present, arranged and presented in a statistical manner, is known as the particle size distribution. The most common way of presenting such distributions as well as the types of distributions important to particles technology, can be found in the specialized literature [21, 22] (see → Particle Size Analysis and Characterization of a Classification Process).

In solid–liquid separation it is important to know the particle size distribution in order to identify which part of it will be separated, and transform this quantity in a measure of efficiency (→ Particle Size Analysis and Characterization of a Classification Process, Section 2.1). The common way of presenting particle size data for solid–liquid separation purposes is in form of a plot. The most common ones are equivalent sphere diameters, equivalent circle diameter, and statistical diameters. The equivalent sphere diameter is the diameter of a sphere which has the same property as the particle itself. Such a property could be the settling velocity. A diameter derived from the settling velocity is known as Stokes diameter and is a very useful quantity for solid–liquid separation, especially to those techniques in which the particle motion relative to the fluid is the governing mechanism.

The amount of particle matter which belongs to specified size classes on the particle axis may be also represented in several ways. Number of particles and mass of particles are the most commons ones, but surface area and volume are used as well. For solid–liquid separation techniques, the most convenient form of expressing amount of particle matter is by mass, because the balances necessary to define performance are normally mass balances.

In general, particle size distributions can be presented as frequencies, *f(x)*, or cumulative frequencies, *F(x)*, which are related to each other by following equation:

$$f(x) = \frac{dF(x)}{dx} \tag{35}$$

the graphical representation of a particle size distribution is usually plotted in a cumulative form. In a typical cumulative plot, points are entered showing the amount of particulate material contributed by particles below or above a specified size. Hence, the curve presents a continuously rising or deceasing character. These oversize and undersize distributions, as illustrated in Figure 5, are simply related by:

$$F(x)_{\text{oversize}} = 1 - F(x) \tag{36}$$

where *F(x)* is the cumulative fraction undersize.

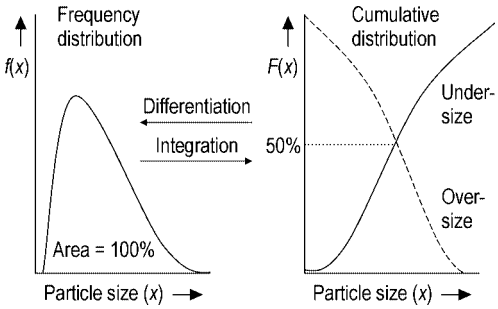

Figure 5. Relationship between frequency and cumulative distributions

A cumulative plot will, therefore, include a broad range of particle sizes. It is often convenient, however, to refer to a single characteristic size for system. Many characteristic sizes have been proposed, most of them involving a mathematical formula. An important one, which can be read off any cumulative plot of the particle size data, is the median particle size. It is defined as that particle size for which the particle amount equals 50 % of the total. If the particle size is represented by number, such point is called the number median size. If mass is used as the measure of particle amount, this parameter is known as the mass median size. The distribution between number and mass median is very important, since they differ generally by a considerable amount. Such difference means that number and mass cumulative plots do not agree for the same system of particles. The weight of a particle, which varies as the cube of its diameter, accounts for this mentioned disagreement.

For practical purposes, it is reasonable to fit an analytical function to experimental particle size distribution data, and then handle this function mathematically in further treatment. It is, for example, very much easier to evaluate mean sizes from analytical functions from experimental data. Several different distribution functions can be found in the literature (→ Particle Size Analysis and Characterization of a Classification Process, Section 2.3). An exhaustive review of some of them has been presented [23].

The *normal distribution* is widely used in most fields of science for the representation of populations. The normal distribution function is given by:

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right] \tag{37}$$

where y is the probability density, x the diameter of the particle, a the arithmetic mean, and σ the standard deviation. From the normal function, if the arithmetic mean a is zero, the probability of occurrence within the interval form the mean (zero) to the value x, i.e., $F(x)$ is given by the *error function*:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^x \exp\left[-\frac{x^2}{2\sigma^2}\right] dx \tag{38}$$

where σ and x have the same meaning as described in the normal function.

Materials with a normal distribution of particle size are relatively rare and are found chiefly among the particulates produced by chemical processes like condensation or precipitation. The importance of this function, however, is that it provides an idealized error distribution built upon the assumption that elementary errors or small causes combine at random to produce the observed effect.

The *log-normal distribution* is probably the most widely used type of function; it is again a two-parameter function, it is skewed to the right and it gives equal probability to ratios of sizes rather than to size differences, as in the normal distribution. Furthermore, the log-normal function is the most useful one among the different types of functions. It can be expressed in the following form:

$$f(x) = \frac{1}{x \ln\sigma_g \sqrt{2\pi}} \exp\left[-\frac{(\ln x - \ln x_g)^2}{2\ln^2\sigma_g}\right] \tag{39}$$

where $f(x)$ is the size distribution function for particle size x, x_g is the geometric mean of the distribution, and σ_g the geometric standard deviation of $\ln x$.

The log-normal distribution has the important advantage of ease of conversion from one type of size distribution to another. It can be shown mathematically that all four particle size distributions, by number, length, surface, and volume (mass), when plotted on log-probability paper are parallel lines with equal linear spacing.

The *Rosin–Rammler distribution* function is another equation widely used in particle size measurement. It is a two-parameter function, usually given as cumulative percentage oversize:

$$F(x) = \exp\left[-\left(\frac{x}{x_r}\right)^n\right] \tag{40}$$

where x_r is a constant giving a measure of the particle size range present and n is another constant characteristic of the material under analysis and gives a measure of the steepness of the cumulative curve. The frequency distribution is obtained by differentiation of Equation 40, which can be reduced to:

$$\log\left[\ln\frac{1}{F(x)}\right] = n\log x - n\log x_r \tag{41}$$

which gives a straight line if $\log(\ln 1/F/(x))$ is plotted against $\ln x$. This is the basis for the Rosin–Rammler graph paper because it is the size corresponding to $100/e = 36.8\%$ and n is the slope of the line. Rosin–Rammler charts from German sources contain edge scales, which with the aid of a parallel rule allow direct estimation of n. For accurate work, the use of large sizes (32.4×25.4 cm) charts designed by Harris [24] is recommended. The various mean sizes available may be easily calculated from the Rosin–Rammler equation; they all involve mathematical tables.

The *Harris model* is a three-parameter equation, which is very versatile and fits many empirical size distributions. Harris [24] showed that most of the widely used two-parameter equations are in fact special cases of the equation:

$$F(x) = \left[1 - \left(\frac{x}{x_o}\right)^s\right]^r \tag{42}$$

where *F(x)* is cumulative percentage oversize, x_o is the maximum size in the sample, s is a parameter concerned with the slope of the log–ln plot in the fine region and r is a parameter concerned with the shape of the log–ln plot in the coarse region.

The ways for measuring particle size are diverse and have been largely improved in recent years. Sieving and sedimentation are the traditional ones, but microscopy, laser diffraction, and stream scanning techniques are currently spreading and finding applications. Several methods for analytical determination of particle size analysis have been reviewed [23]. Although the most novel techniques have the advantage of quick response, they have proved to be very selective. Therefore, most of them are of quite limited application. A rule of thumb should be to use the analytical technique whose indirect measurement agrees, as much as possible, with the specific application. For example, in solid–liquid separations one is concerned with particles settling by gravitational forces, so techniques using this sort of forces to evaluate particle size distribution would be most appropriate.

To evaluate the efficiency of separation it is necessary to take into account that solid–liquid separation is basically an imperfect process. Whereas the underflow is always wet slurry, the overflow can be considered as a turbid liquid. This general behavior is critic in hydrocyclones, as they always, despite the presence of solids in the liquid, will split the flow into two streams. Therefore, the aim of operation of a hydrocyclone will be not only to obtain the above-mentioned slurry, but also to deliver is as free from liquid as possible. Equally, for clarification purposes, the overflow should be drawn from the hydrocyclone with the least possible concentration of solids. As Tenberger and Rietema [25] have stated, a single efficiency number can never be capable of fully describe the result of separation, except when it is ideal.

The imperfection of solid–liquid separation has caused the need to express efficiency by different means. A good review of these mentioned efficiencies is presented in [26].

The first and most obvious definition of separation efficiency is simply the overall mass recovery as a fraction of the feed flow rate. According to Figure 6:

$$E_t = \frac{M_c}{M} \tag{43}$$

where all the components are as defined in Figure 6.

If there is no accumulation of solids in the separator then:

$$M = M_c + M_f \tag{44}$$

and there is a choice of three possible combinations of the material streams for the total efficiency testing. It can be shown [3] that if all the operating conditions are equal, the most accurate estimation of the local efficiency comes from the two leaving streams.

As has been already mentioned, hydrocyclones act as flow dividers. Thus, they also split the solids in at least the same ratio, as is the ratio of underflow to throughput, R_f:

$$R_f = \frac{U}{Q} \tag{45}$$

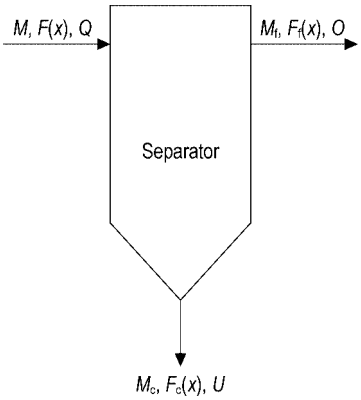

Figure 6. Schematic diagram of a separator. M: mass flow rate of solids in the feed, M_c: mass flow rate of separated solids, M_f: mass flow rate of nonseparated solids, $F(x)$: cumulative percentage oversize of feed solids, $F_c(x)$: cumulative percentage oversize of separated solids, $F_f(x)$: cumulative percentage oversize of nonseparated solids, Q: volumetric flow rate of feed suspension, U: volumetric flow rate of underflow suspension, O: volumetric flow rate of overflow suspension

The total efficiency defined in Equation 43 will thus include the effect of this flow splitting known as dead flux. For this reason, a better way of expressing efficiency of separators such as hydrocyclones, is by means of a reduced efficiency which has been defined as:

$$E\prime_t = \frac{E_t - R_f}{1 - R_f} \tag{46}$$

If complete separation is achieved, providing that there is no accumulation of solids, the value of the total efficiency from Equation 43 will be the unit, and so will be the value of the reduced total efficiency. If, on the other hand, there is no separation at all, E_t will equal R_f in order to satisfy the requirements of an efficiency definition, leading to a value of zero in Equation 46.

The total efficiency defined by Equation 43 includes all particle sizes present in the feed solids. If only a narrow range of particle sizes is of interest, another efficiency of separation particular to that range can be defined. A mathematical expression of such partial efficiency is:

$$E_p = \left[\frac{M_c}{M}\right]_{x_1/x_2} \tag{47}$$

where x_1 and x_2 represent the particle size limits of a definite range.

If the particle size range in Equation 47 becomes infinitesimal, the obtained efficiency corresponds to a single particle size x and it is known as the grade efficiency, defined by:

$$G(x) = \left[\frac{M_c}{M}\right]_x \tag{48}$$

The grade efficiency has become a very useful definition, because most industrial powders consist of an infinite number of differently sized particles. Thus, a single particle size really corresponds to a range of particles having almost similar sizes. Therefore, the grade efficiency of most separation equipments is a continuous function of x. This function is seldom expressed analytically but graphically. An S-shaped curve is usually obtained for separators like hydrocyclones, in which inertial or gravity body forces perform the separation.

As the value of the grade efficiency has the character of probability, plotting the probability for any given size fraction against particle size would give a curve as known in Figure 7. This sort of curve is normally called a grade efficiency curve.

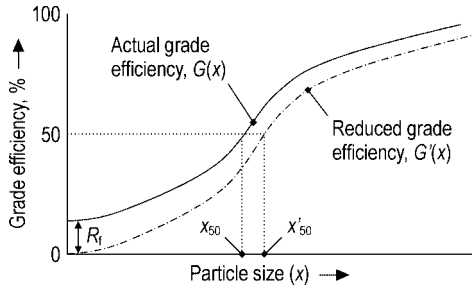

Figure 7. Grade efficiency and reduced grade efficiency curve for hydrocyclones

A grade efficiency curve for hydrocyclone is derived from screen analysis data on the feed, overflow and underflow streams. As mentioned before, it is practically impossible to fraction common systems of particles in a finite number of sizes. It can be shown [27] that a practical grade efficiency curve is really derived from a stepwise calculation, drawing a line through the midpoints of size intervals. A curve thus derived does not pass through the origin. This can be explained bearing in mind that, as has been previously emphasized, a hydrocyclone is a flow divider, so the underflow always contains a certain quantity of very fine particles which simply follow the flow, and are split in the same ratio as the liquid. The apparently finite efficiency for

fine particles is, therefore, equal to the underflow to throughput ratio R_f as shown in Figure 7. A reduced grade efficiency, similar to the reduced efficiency represented by Equation 46 can then be obtained as:

$$G'(x) = \frac{G(x) - R_f}{1 - R_f} \tag{49}$$

As can be seen in Figure 7, the correspondent curve to this definition does cross the origin.

Hydrocyclones are among those equipments whose separation performance is highly dependent on particle size. Thus, according to this, an overall efficiency of separation is not the best way to evaluate their performance. For this reason, it is advisable to express their efficiency for each size of particle in the range measured. As experience has shown, the only single number, which in some way gives the separation power of a hydrocyclone, is the *cut size* x_{50} (for a definition of cut size see → Particle Size Analysis and Characterization of a Classification Process, Section 3.1.4). All particles above the cut size will be generally going to the underflow, whereas those below the cut size will be normally reported to the overflow.

As can be concluded, the cut size is derived from a grade efficiency curve. As the calculation and plotting of a full grade efficiency curve is a time-consuming task, alternative ways of determining the cut size have been proposed [28, 29].

The *sharpness of separation* can be related to the shape of the curve in a number of manners. Knowing the grade efficiency curve, comparisons of classification sharpness of different equipments can be made by plotting *G(x)* against the dimensionless size x/x_{50}, and comparing the resultant normalized curves. To define the curve steepness, a ratio of two sizes corresponding to two different percentages on the grade efficiency curve on either side, and 50 % equidistant, can be used. This parameter is called the *sharpness index* and is represented by:

$$H_{25/75} = \left[\frac{x_{25}}{x_{75}}\right] \tag{50}$$

It was previously stated that the grade efficiency curve does not go through the origin. The cut size as has been defined is derived from this curve. If, in order to practically assess the performance of hydrocyclones, a reduced grade efficiency curve is used, the particle size which gives a 50 % efficiency in such a curve is called the reduced cut size and is represented by x'_{50}.

The maximum obtainable efficiency related to particle size would be that minimum particle size with 100 % probability of being reported to the underflow. Graphically, by extrapolating the end part of the curve to the horizontal axis, such size will be obtained. It has been proved that in practice, the maximum of the efficiency is around 98 %, and the minimum size correspondent to this efficiency is represented by x_{98} and known as the approximate limit of separation.

Several correlations have been proposed in literature to evaluate the cut size. They are either empirical or theoretical, and are expressed in the traditional British System of units, as well as in SI units, even in both. Perhaps the most important criterion for classifying the proposed correlation for cut size, is by dividing them into two groups considering whether the feed concentration has been taken into account. The main correlations, which do not consider the feed concentration as a variable, will be reviewed as follows.

Bradley [2] proposed:

$$x_{50} = \left[\frac{18\pi\mu(1-R_f)}{16LQ(\rho_s-\rho)}\right]^{1/2} \left[\frac{2.3D_o}{D_c}\right]^n \left[\frac{D_i^2}{\alpha}\right] \tag{51}$$

where Q is the flow rate, D_o, D_c and D_i are the dimensions defined in Figure 2, and (and n are empirical constants depending on the cyclone design and the fluid properties.

Rietema [30] derived a dimensionless, characteristic cyclone number C_{y50} represented by:

$$C_{y50} = x_{50}^2\,(\rho_s - \rho)\,\frac{L\Delta P}{\mu\rho Q} \tag{52}$$

where ΔP is the pressure drop. C_{y50} is determined experimentally and is not directly related to the more important hydrocyclone variables. However, Rietema has demonstrated the usefulness of the parameter and quotes a value of 3.50 for optimal separation, corresponding to the conditions: $L/D_c = 5$, $D_i/D_c = 0.28$, $D_o/D_c = 0.34$ and $l/D_c = 0.4$, where all the dimensions are as defined in Figure 2.

Dahlstrom [31] proposed:

$$x_{50} = \frac{C(D_o D_i)^{0.68}}{Q^{0.53}} \left[\frac{1.73}{\rho_s - \rho}\right]^{1/2} \tag{53}$$

where C is a constant sensitive only to major changes in the cyclone dimensions. This equation holds well for cyclones in which the length

of the cylindrical section equals that of the cone, and has been shown to hold for diameters of 3 to 14 inches (7.5 to 35 cm), for flows containing up to 20 % solids by weight.

Yoshioka and Hotta [32] developed an equation based on the orbital concept and equilibrium cone surface defined by the end of the vortex finder and the cone apex. The equation is:

$$x_{50} = 0.2 D_c^{0.1} D_i^{0.6} D_o^{0.8} \left[\frac{\mu}{Q(\rho_s - \rho)} \right]^{1/2} \tag{54}$$

which is expressed in SI units; the constant 0.2 was experimentally evaluated.

De Gelder [33] has derived an equation, which does not include experimental values:

$$x_0 = D_o \left[\frac{0.349 Re(D_c - kD_o)}{D_i} \right] \left[\frac{\rho_s}{\rho} - 1 \right]^{1/2} \tag{55}$$

where *Re* is the Reynolds number, x_0 is the diameter of particle which is just not collected, i.e., with centrifugal efficiency of 0 % and *k* is a constant, function of $D_i^2/D_o D_c$. Although Equation 55 was first developed for gas cyclones, it is said to hold well for hydrocyclones.

Lynch and Rao [34] presented the following empirical correlation:

$$\log x\prime_{50} = \frac{D_o}{2.6} - \frac{D_u}{3.5} + \frac{\Delta P}{10.7} - \frac{Ql_o}{52} + K \tag{56}$$

where x_{50}^* is in micrometers, D_o and D_u are in inches, ΔP is in pounds per square inch (psi), Ql_o (mass flow rate of the liquid in the overflow) is in ton/h, and *K* is a function of the material and the hydrocyclone (2.0 for silica and 2.5 for copper ore).

3.4.2. Capacity, Pressure Drop

The capacity of a hydrocyclone can be evaluated in terms of the volume flow rate delivered, also known as throughput, which depends on the available pressure drop. Thus, there is a loss of pressure across the unit. The pressure drop is an important variable and is more easily defined than efficiency. Relationships between Q and ΔP can be developed by means of established theory for the flow of liquids in pipes. There are a number of pressure drop correlations reported in the literature, some of which will be described.

Trawinski [35] proposed the correlation:

$$Q = K D_i D_o \left[\frac{\Delta P_g}{\rho} \right]^{1/2} \tag{57}$$

where *K* is a factor, which contains diameter ratios, fraction loss, and cone angle variables. For $\theta = 15° - 30°$, *K* equals 0.5.

Bradley's approach [2] leads to the following expression:

$$\frac{(\Delta P/\rho)}{(v_i^2/2g)} = \frac{\alpha^2}{n} \left[(D_c/D_o)^{2n} - 1 \right] \tag{58}$$

where the left-hand side group is the pressure difference in terms of inlet velocity heads and α is the inlet velocity loss coefficient v_c/v_i whereas *n* is the power from Equation 23.

4. Effect of Variables on Hydrocyclone Performance

A single conclusion can be made from the previously survey: the operation of hydrocyclones is quite complex. There is not a unified theory to fully explain the mechanisms of flow and particle separation inside a hydrocyclone. It is normally accepted that the overall behavior of the particles of solids inside a unit will depend on factors contributed by the liquid, by the solid, and by the apparatus design. These factors may be interrelated or independent of each other. In order to study the effect of such factors upon the performance of hydrocyclone units, they have been usually divided into operational variables and design variables.

4.1. Operation Variables

First of all, a pressure drop must exist in order to drive the feed suspension across the unit. An increment in this pressure drop will decrease the cut size as well as increase the sharpness of separation and flow rate. According to this, great mass recoveries could be obtained raising the pressure drop. Although such fact is literally true, it is normally an uneconomical measure because large pressure loss increases are necessary. As has been mentioned earlier, practical values of pressure drop range between 0.34 and 6 bar. However, for small cyclones, Trawinski [27] has suggested an upper limit of 4 bar. As in other separation techniques, a finite density

difference between the phases involved is essentially necessary. Without it, no separation can be achieved. As expected, experience has shown an increase in separation efficiency as the density difference increases.

The effect of feed solids concentration on separation performance is well recognized in practice. Generally, increasing input concentrations reduces the total efficiency. Svarovsky and Marasinghe [36] have proposed a semiempirical correlation, which allows scale-up prediction of performance with feed concentrations higher than 8 vol %. Neesse et al. [37] compared dilute flow (low solids concentration of the feed) and dense flow (high solids concentration of the feed) separations, focusing their study on the effect of turbulence.

Perhaps the least understood effect on performance of a hydrocyclone is that of the viscosity. Very little quantitative information exists on the literature about it. It is normally accepted that viscosity has little effect on separation efficiency. From simple theory it can be proved that medium viscosity will only have an effect at low Reynolds numbers (i.e., laminar flow), but because turbulence is promoted by pumping, the effect of viscosity will only be appreciable when the flow is forced by static head or in very large units. Nappier-Munn [38] found concordance with theory about viscosity effects, while Williamson et al. [39] studied the effect of viscosity, along with pressure drop, in reducing the minimum cut size of particles that can be separated effectively. They found that cut size falls as the pressure drop rises and as viscosity is reduced.

The effect of less common variables has also been studied. Fontein et al. [40] examined the influence of some uncommon parameters such as the roughness of the cyclone wall, as well as the shape of the solid particles on suspension. They found that rough-walled cyclones show lower efficiencies than smooth-walled ones. According to them, this is mainly due to the fact the rotation flow is greatly retarded by the rough wall, thus decreasing the centrifugal force, which affects separation.

4.2. Design Variables

Some modifications and recommendations about the geometry of hydrocyclones have appeared in the literature, mainly from manufactures. Despite many variations in design and suggestions for varied configurations, the so-called standard cyclones, such as the Rietema cyclone, have proved the most efficient in separation and capacity.

The most common materials of construction are steel, aluminum oxide, ceramics, polyurethane, and other plastics. Sometimes the inner body is lined with hard materials in order to minimize wear caused by operation. Soft rubber is popular and suitable where temperatures do not exceed 60°C and where hydrocarbon oils or solvents are absent. Ceramic liners can also be used. Wear-resistant metals are useful when particle sizes are above 100 μm or where tramp material is suspected, which could damage rubber linings. With temperatures above 60 °C or in presence of hydrocarbons, neoprene or nitrile, as well as polyurethane rubbers have been successfully employed.

The effects of some geometry variables on hydrocyclone performance are shown in Table 1.

Table 1. Effects of some design variables on hydrocyclone performance

Variable	Effect
Cyclone diameter increases	cut size increases, pressure drop usually decreases
Cyclone inlet diameter increases	gravity force in cyclone decreases, cut size increases, capacity falls, pressure drop decreases
Overflow diameter increases	cut size increases, risk of coarse sizes evident
Underflow diameter increases	brings excess fines from liquid phase into underflow
Cyclone shape becomes longer	decreases cut size sharpness separation
Cyclone wall become rougher	capacity increases, efficiency decreases

5. Selection and Design of Hydrocyclone Systems

Hydrocyclones differ from other solid–liquid separation equipments as they have a broad number of variables involved in their operation. The influence of some variables is not completely understood, and their effect may be completely

different from one unit to another. Theses are some of the reasons why the task of selecting a hydrocyclone system is difficult to deal with. Generalizations of performance characteristics can seldom be made, and the choice necessarily relies upon semiempirical relationships or models, which are, inevitably, of limited application. In selecting or designing hydrocyclones for a specific duty, the recommended way is either using manufactures catalogues or to make calculations to obtain cut point and throughputs.

5.1. Analytical Solutions

The previously cited formulae for the computation of cut points and capacities can be employed as a first approximation. The actual choice can be mainly based on the function to be performed by the system. In classification applications, for instance, the cut size is usually fixed and the design is simply based upon achieving that cut size. On the other hand, in separation, a certain degree of efficiency is normally required, and the resulting design must be a compromise between technical and economical considerations. The adaptability to industrial scale can represent a real problem. Moir [41] pointed out that some of the correlations for calculating cut points and throughputs for cyclones might not be suitable for industrial problems, since they were developed using slurries with low solids contents and small units. An alternative to high capacities is use of multiple hydrocyclone systems for industrial production. Gerrard and Liddle [42] have described two methods for selection of units to be used in parallel for a given duty. Many other authors have mentioned the advantage of using multicyclone systems.

5.2. Graphical Solutions

A series of charts or nomographs have been also developed for the estimation of effects of changes in operating variables on performance of hydrocyclones. Zanker [43] presented a series of equations incorporated into some nomographs in order to make trial-and-error calculations faster. Rao and Nageswararao [44] simplified the use of charts to just one nomograph. Again, combined with some calculations, they claimed such a graph to be a useful tool in selecting hydrocyclones for specific duties. Another proposed graphical method is given in [45]. The authors emphasized, however, that application of the technique should be for guidance purposes only.

5.3. Manufactures' Choice

Several equipment manufactures have published literature about selection of hydrocyclones in the form of brochures, technical reports, and so on. In many International Conferences held around the world some manufactures present papers about selection, innovations, etc., as well as table-top exhibitions of their available equipment.

Table 2. Summary of some known hydrocyclone designs

Cyclone type and size of hydrocyclone	D_i/D_c	D_o/D_c	l/D_c	L/D_c	θ	$Stk_{50}Eu^{**}$	$K_p{}^{**}$	$n_p{}^{**}$
Rietema's design* D_c = 0.075 m	0.280	0.340	0.40	5.00	20°	0.0611	24.38	0.3748
Bradley's designD_c = 0.038 m	0.133	0.200	0.33	6.85	9°	0.1111	446.5	0.3230
Mozley cyclone 1D_c = 0.022 m	0.154	0.214	0.57	7.43	6°	0.1203	6381	0
Mozley cyclone 2D_c = 0.044 m	0.160	0.250	0.57	7.71	6°	0.1508	4451	0
Mozley cyclone 3D_c = 0.022 m	0.197	0.320	0.57	7.71	6°	0.2182	3441	0
Warman 3″ model RD_c = 0.076 m	0.290	0.200	0.31	4.00	15°	0.1709	2.618	0.8000
RW 2515 (AKW)D_c = 0.125 m	0.200	0.320	0.80	6.24	15°	0.1642	2458	0

* Optimum separation.
** Optimum separation.

5.4. Dimensionless Scale-Up of Hydrocyclones

5.4.1. Introduction

Because of the hydrocyclones' inflexibility, the methods of selection just reviewed can be only used as an approximation. The correlations proposed are often derived from studies of effects of geometrical proportions on capacity and separation efficiency. In this sense, if literature on hydrocyclones is rigidly followed when designing a unit, it might lead to a unique hydrocyclone choice, whose geometry could has been never tested before, so reliable prediction of its performance may be misleading. Based on this and many other facts and factors, a scale-up procedure to select hydrocyclones, attributed mainly to SVAROVSKY [1] has been proposed. The method consists of selecting a known design of hydrocyclone and, by using a series of dimensionless relations, making the necessary predictions as changes on fixed variables are introduced. A summary of some known hydrocyclone designs is presented in Table 2.

As the performance characteristics of hydrocyclones involve a great number of variables, the use of dimensionless groups in this method is an obvious advantage. The basic concepts underlying dimensionless analysis and scale-up procedures have been used for long time in classical mechanics principally. Perhaps the most common method of dimensional analysis is that attributed to LORD RAYLEIGH known as the Pi theorem. The basic steps of dimensional analysis can be found in (→ Scale-Up in Chemical Engineering) [43].

5.4.2. Definition and Derivation of Dimensionless Groups for Hydrocyclones

In the case of hydrocyclones, selection and operation is based on the relationships between the pressure drop and flow rate and the relationship between separation efficiency and flow rate. The pressure drop versus volumetric flow rate relationship is usually expressed as $Eu = f\,(\mathrm{Re})$, where *Eu* is the Euler number. These dimensionless groups are graphically related as shown in Figure 8. The Euler number is in fact a pressure loss factor, easily defined as the limit on the maximum characteristic velocity v obtained by a certain pressure drop ΔP across the hydrocyclone. It can be expressed as:

$$Eu = \frac{2\Delta P}{\rho v^2} \tag{59}$$

where ρ is the density of the fluid.

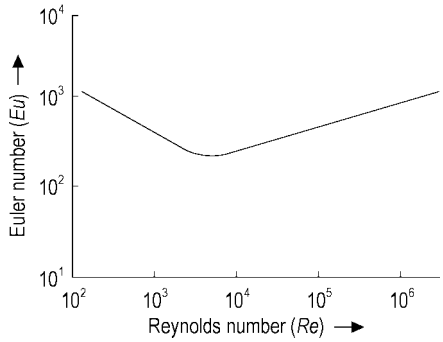

Figure 8. A typical plot of *Eu* versus *Re* for hydrocyclones

The Reynolds number defines flow characteristic of the system and in the case of hydrocyclones, the characteristic dimension may be taken as the cyclone body diameter D_c, i.e.:

$$Re = \frac{D_c v \rho}{\mu} \tag{60}$$

where μ is the medium viscosity.

The relationship between separation efficiency and flow rate is not significantly influenced by operational variables, so it is commonly expressed in terms of cut size x_{50}. The use of the cut size to define efficiency of hydrocyclone is of the utmost importance since, as has been previously emphasized their performance is highly dependent on particle size. Since cut size implies size of particles likely to be separated, it follows that such particles must be influenced by forces exercised on the suspension. The forces developed in a hydrocyclone can be analyzed by sedimentation theory, and a dimensionless group thus derived, the Stokes number, will include cut size.

The Stokes number is a very useful theoretical tool and, for the case of hydrocyclones, its derivation may be carried out as follows. Considering a settling tank of length L and wide W, it would have a settling area of $A = LW$, and if the depth is H, then the retention volume is $V =$

AH. With a flow through the tank of Q (m^3/h), the detention time is given by:

$$t_{\mathrm{d}} = \frac{V}{Q} \tag{61}$$

Particles to be removed must settle, under gravity influence, through the depth H requiring t_{s}, where:

$$t_{\mathrm{s}} = \frac{H}{u_{\mathrm{t}}} \tag{62}$$

u_{t} being the settling rate defined by Equation 6.

Only particles for which t_{s} is equal or less than t_{d} will be removed. Hence, by equating t_{s} and t_{d} a cut point or separation mesh is defined as:

$$\frac{H}{u_{\mathrm{t}}} = \frac{V}{Q} \tag{63}$$

transposing for u_{t}:

$$u_{\mathrm{t}} = \frac{HQ}{V} = \frac{Q}{A} \tag{64}$$

If Equation 64 is then equated to Equation 6, the limiting size of particles to be settled can be derived giving:

$$x^2 = \frac{18\mu Q}{g\left(\rho_{\mathrm{s}} - \rho\right) A} \tag{65}$$

For centrifugal systems the same replacement done in Equation 6 leads to:

$$x^2 = \frac{18\mu Q}{R\omega^2\left(\rho_{\mathrm{s}} - \rho\right) A} \tag{66}$$

Since $\omega = v_t/R$, where v_{t} is the tangential velocity of the particle, Equation 66 can also be expressed as:

$$x^2 = \frac{18\mu QR}{v_t^2\left(\rho_{\mathrm{s}} - \rho\right) A} \tag{67}$$

The radial settling velocity in a hydrocyclone is due to the centrifugal acceleration, which is proportional to the square of the tangential velocity of the particle and indirectly proportional to the radius of the particle position. As the tangential motion of the particle is unopposed, the tangential particle velocity can be taken as equal to the tangential component of the fluid velocity at the same point. For the same flow regime, the velocities anywhere in the flow in a hydrocyclone are proportional to a characteristic velocity v while the position radii are proportional to the cyclone diameter D_{c}. Under such assumptions Equation 67 can be approximated to:

$$x^2 = \frac{18\mu Q D_{\mathrm{c}}}{v^2\left(\rho_{\mathrm{s}} - \rho\right) A} \tag{68}$$

Considering the superficial velocity in the hydrocyclone body as the characteristic velocity, the volumetric flow rate equals $\left(v\pi D_c^2\right)/4$, also $A = \pi D_c^2/4$ for the cyclone body. Thus, substituting Q and A into Equation 68:

$$x^2 = \frac{18\mu D_{\mathrm{c}}}{v\left(\rho_{\mathrm{s}} - \rho\right)} \tag{69}$$

Rearranging Equation 69 to express it in terms of inertial forces and hydrodynamic forces; the dimensionless group known as Stokes number is obtained, i.e.:

$$Stk = \frac{x^2\left(\rho_{\mathrm{s}} - \rho\right) v}{18\mu D_{\mathrm{c}}} \tag{70}$$

Furthermore, if the dimension x is replaced by the specific cut size, x_{50}:

$$Stk_{50} = \frac{x_{50}^2\left(\rho_{\mathrm{s}} - \rho\right) v}{18\mu D_{\mathrm{c}}} \tag{71}$$

or:

$$Stk_{50}(r) = \frac{x_{50}'^2\left(\rho_{\mathrm{s}} - \rho\right) v}{18\mu D_{\mathrm{c}}} \tag{72}$$

if the reduced cut size x_{50}' or $x_{50}(r)$ is incorporated.

5.4.3. Scale-Up at Low Concentrations

At low concentrations of the feed (less than 1 vol %), the flow pattern in a hydrocyclone is not affected by the presence of particles in the flow and particle—particle interaction is negligible. As only few particles report to the underflow, the underflow-to-throughput ratio can be assumed to have no effect on the cut size. If this is the case, the dimensional analysis gives two basic relations between the three above described dimensionless groups:

$$Stk_{50}Eu = \text{constant} \tag{73}$$

$$Eu = K_{\mathrm{p}} Re^{n_{\mathrm{p}}} \tag{74}$$

where K_{p} and n_{p} are empirical constants for a family of geometrically similar cyclones. The product $Stk_{50}Eu$ depends on the cyclone design, and on the underflow to throughput ratio. Its value range from 0.06 to 0.33. The value of n_{p} is usually between 0.0 and 0.4 [1].

5.4.4. Scale-Up at High Concentrations

At higher concentrations, the feed concentration as a fraction of volume C has to be included as an additional dimensionless group. The feed concentration of solids is a very important variable, which affects the separation efficiency and the pressure drop of a given hydrocyclone. It has been already mentioned the decreasing in efficiency as feed concentration increases. This means that the underflow-to-throughput ratio has to be increased (to allow greater volume of separated solids to be discharged). For this reason, most hydrocyclones are equipped with either a continuously variable underflow orifice, or a series of replaceable nozzles.

Theoretical and empirical approaches of the effect of feed solids concentration in hydrocyclones have appeared in the literature. Svarovsky and Marasinghe [36] have developed the following expression for the effect of high-feed solids concentration:

$$Stk_{50}(r) = k_1\,(1-R_{\mathrm{f}})\exp(k_2 C) \tag{75}$$

where $Stk_{50}(r)$ includes the reduced cut size, as already defined, while k_1 and k_2 are empirical constants, which depend on the hydrocyclone configuration as well as on solids properties. The correlation has proved to hold well for concentrations above 8 vol %, and the values of the constants k_1 and k_2 were found to be 9.05 $\times 10^{-5}$ and 6.461, respectively, for limestone and a hydrocyclone manufactured by Amberger Kaolinwerke (AKW).

5.4.5. Considerations for non-Newtonian Behavior

When fine powders are dispersed into liquids the resultant suspension is usually non-Newtonian. The non-Newtonian behavior of suspensions is normally a direct function of concentration. In this sense, the viscosity of a dilute suspension nearly equals that of the suspending liquid. Thus, when dealing with suspension of low solids contents either the apparent suspension viscosity or the suspending liquid viscosity can be used for characterization purposes. However, for more concentrated suspensions non-Newtonian behavior may be present and their viscosities cannot be considered constant anymore; their flow behavior must, therefore, be evaluated by other means.

The point above mentioned, make necessary to include parameters that characterize non-Newtonian fluids in the previously developed dimensionless groups, if they are going to be used in scale-up of hydrocyclones dealing with suspensions, which might be non-Newtonian. In this case, an approach similar to that employed in the derivation of Equation 17 may be carried out. Firstly, in this type of fluid, the drag coefficient, C_{D}, for steady-state creeping flow past a sphere based on a macroscopic momentum and energy balance of Wasserman and Slattery [47] is:

$$C_{\mathrm{D}} = \frac{24}{Re*} \tag{76}$$

where $Re*$ is a generalized Reynolds number analogous to that defined by Equation 17, being its characteristic dimension x, i.e,. the particle diameter, and $\gamma = K'(3)^{n'-1}$. Also, $K' = K[(2n'+1)/3n']^{n'}$, for this case.

For Newtonian fluids, an approximation of the drag coefficient in Stokes and the intermediate region [48] is:

$$C_{\mathrm{D}} = \left[\tfrac{24}{Re}\left(\tfrac{24}{Re}+4.5\right)\right]^{1/2} \qquad (Re<500) \tag{77}$$

a modification of Equation 77 to better fit the data is:

$$C_{\mathrm{D}} = \left[\tfrac{24}{Re^{\alpha_1}}\left(\tfrac{24}{Re^{\alpha_2}}+\alpha_3\right)\right]^{1/2} \qquad (Re*<500) \tag{78}$$

where the alpha constants are $\alpha_1 = 1.1$, $\alpha_2 = 0.2$, and $\alpha_3 = 7.5$.

By substitution of $24/Re*$ for $24/Re$ in Equation 78 and modification of the constants to ensure a friction factor of 0.44 at $Re* = 500$, the resultant non-Newtonian drag coefficient correlation is:

$$C_{\mathrm{D}} = \left[\tfrac{24}{Re*^{\alpha_1}}\left(\tfrac{24}{Re*^{\alpha_2}}+\alpha_3\right)\right]^{1/2} \qquad (\mathrm{Re}*<500) \tag{79}$$

where the alpha constant values are as previously noted.

The terminal settling velocity for non-Newtonian fluids $(Re<500)$ is found by equating Relation (77) with the generalized form of the drag coefficient, i.e.:

$$\frac{4gx}{3u_t^2}\left[\frac{\rho_{\mathrm{s}}-\rho}{\rho}\right] = \left[\frac{24}{Re*^{\alpha_1}}\left(\frac{24}{Re*^{\alpha_2}}+\alpha_3\right)\right]^{1/2} \tag{80}$$

Substituting Equation 17 and eliminating Reynolds number, Equation 80 becomes:

$$\left[\frac{gx^{n'+1}\left(\rho_s-\rho\right)}{18\gamma}\right]^2 = u_t^{2n'}+\frac{\alpha_3}{24}\left[\frac{\rho x^{n'}}{\gamma}\right]^{\alpha_2} u_t^{2n'+\alpha_2(2-n')} \quad (81)$$

For small Reynolds numbers $(\mathrm{Re}*<0.1)$, the second term of the right-hand side of Equation 81 is negligible and the simplification leads to:

$$u_\mathrm{t} = \left[\frac{gx^{n'+1}(\rho_\mathrm{s}-\rho)}{18\gamma}\right]^{1/n'} \quad (\mathrm{Re}*<0.1) \quad (82)$$

For Newtonian fluids $n' = 1$, γ equals μ, so Equation 82 becomes the familiar expression of Stokes law as described by Equation 6.

Following a similar procedure as when obtaining Stokes' number expression, Equation 82 can be equated to Equation 6, i.e.:

$$\frac{Q}{A} = \left[\frac{gx^{n'+1}\left(\rho_\mathrm{s}-\rho\right)}{18\gamma}\right]^{1/n'} \quad (83)$$

transposing for $x^{n'+1}$:

$$x^{n'+1} = \frac{18\gamma Q^{n'}}{g\left(\rho_\mathrm{s}-\rho\right)A^{n'}} \quad (84)$$

Again, considering centrifugal acceleration:

$$x^{n'+1} = \frac{18\gamma R Q^{n'}}{v_t^2\left(\rho_s-\rho\right)A^{n'}} \quad (85)$$

and, under the assumptions made before:

$$x^{n'+1} = \frac{18\gamma Q^{n'} D_\mathrm{c}}{v^2\left(\rho_\mathrm{s}-\rho\right)A^{n'}} \quad (86)$$

Substituting Q and A and rearranging terms, the Stokes number expression for non-Newtonian systems is obtained, i.e.:

$$Stk* = \frac{x^{n'+1}\left(\rho_\mathrm{s}-\rho\right)v^{2-n'}}{18\gamma D_\mathrm{c}} \quad (87)$$

It can be noticed that, again, that for Newtonian fluids this generalized Stokes number reduces to the common expression represented by Equation 12.

Finally, since concentration is an important factor in this case, replacing the reduced cut size for the simple particle size, x:

$$Stk_{50}^*(r) = \frac{x_{50}^{n'+1}(r)\left(\rho_\mathrm{s}-\rho\right)v^{2-n'}}{18\gamma D_\mathrm{c}} \quad (88)$$

in which $x_{50}(r)$ is preferred to x'_{50} in order to avoid confusion.

5.4.6. Empirical Models

An exhaustive study for concentrations up to 10 vol % was carried out by Medroho [49] in order to prove the applicability of Equations 73, 74 and 75. He employed three geometrically similar hydrocyclones of Rietema's [30] optimum geometry and obtained the following relations:

$$Stk_{50}(r)Eu = 0.047[\ln(1/R_\mathrm{f})]^{0.74}\exp(8.96C) \quad (89)$$

$$Eu = 71(Re)^{-0.116}(D_\mathrm{i}/D_\mathrm{c})^{-1.3}\exp(2.12C) \quad (90)$$

$$R_\mathrm{f} = 1218(D_\mathrm{u}/D_\mathrm{c})^{-4.75}(Eu)^{-0.30} \quad (91)$$

where R_f is the underflow-to-throughput ratio, C is the volume fraction of solids, D_i, D_c, and D_u are the inlet, body, and underflow diameters of the hydrocyclone, respectively. Covering also the low feed concentration range (i.e. up to 10 vol %), Antunes and Medroho [50] developed a set of equations bearing similarity to those presented above, but for hydrocyclones of Bradley's [2] geometry.

Ortega-Rivas and Svarovsky [51] have proposed a model for concentrated suspensions and accounting for non-Newtonian behavior of the treated slurries. They covered a feed concentration range from 5 to 25 vol % and used geometrically similar hydrocyclones following Rietema's [30] proportions. The relationships that they derived are presented below:

$$Stk_{50}^*(r)Eu = 0.006[\ln(1/R_\mathrm{f})]^{2.37}\exp(6.84C) \quad (92)$$

$$Eu = 1686(Re^*)^{-0.035}\exp(-3.39C) \quad (93)$$

$$R_\mathrm{f} = 32.8(D_\mathrm{u}/D_\mathrm{c})^{1.53}(Re^*)^{-0.34}\exp(3.70C) \quad (94)$$

where the dimensionless Reynolds and Stokes numbers include the parameters of characterization of non-Newtonian suspension previously described. Equations 92, 93 and 94 allow reliable hydrocyclones design and scale-up at concentrations up to 25 vol % of shear-thickening suspensions. They apply to Rietema's geometry, but there is some evidence suggesting that they could equally be employed to other geometries, because reasonable agreement has been found with previous models of hydrocyclone design and performance. The authors also claim that

for the case of Newtonian suspensions, i.e., low solids contents, the above mentioned equations would reduce to the common definitions so they conclude their model is of wide applicability.

All the above described relationships have the advantage of presenting graphical functions, such as the one shown in Figure 8. Those graphs can be handled for design purposes by manipulating variables and predict, in a practical manner, operation effects of hydrocyclone technology.

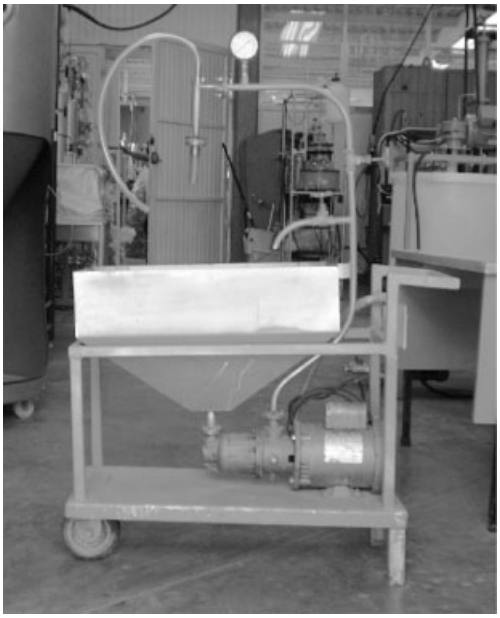

Figure 9. Hydrocyclone test rig

5.4.7. Practical Applications

In order to test the different models previously described in this article, experimental runs in small hydrocyclone rigs are advised. Standard geometries of cyclones, such as the RIETEMA's hydrocyclone, can be used to run tests at different diameters and pressure drops. As previously mentioned, many different design graphs describing the relationships between dimensionless groups may be available for variable manipulation and prediction of best performance variables. A hydrocyclone test rig is shown in Figure 9. The units are changeable and they may be equipped with removable underflow nozzles (Fig. 10.), in order to verify the effect that variations in the underflow orifice diameter may have in the cut size. Once a number of tests have been performed, the possible scale-up to industrial size may follow.

Figure 10. RIETEMA's proportions hydrocyclone with replaceable underflow spigots

Since cut size is a direct function of hydrocyclone diameter, the best efficiency for a number of useful applications would be given by small units. Cut sizes down to 2 μm have been reported in 10 mm diameter hydrocyclones [51]. On the other hand, also capacity is a direct function of hydrocyclone diameter, so larger units will deliver higher flow rates. Therefore, it would be necessary to reach a compromise between efficiency and capacity for industrial operation. In this sense, a basic small geometry may be kept to cut as fine particles as possible, while series arrangements may be required to render industrial capacities. A number of arrangements have been experienced, but probably the most efficient is the radial (spider) arrangement with the hydrocyclones axes horizontally placed like spokes in a wheel lying down. As shown in Figure 11, one "wheel" or hydrocyclone deck accommodates a number of "spokes" or hydrocy-

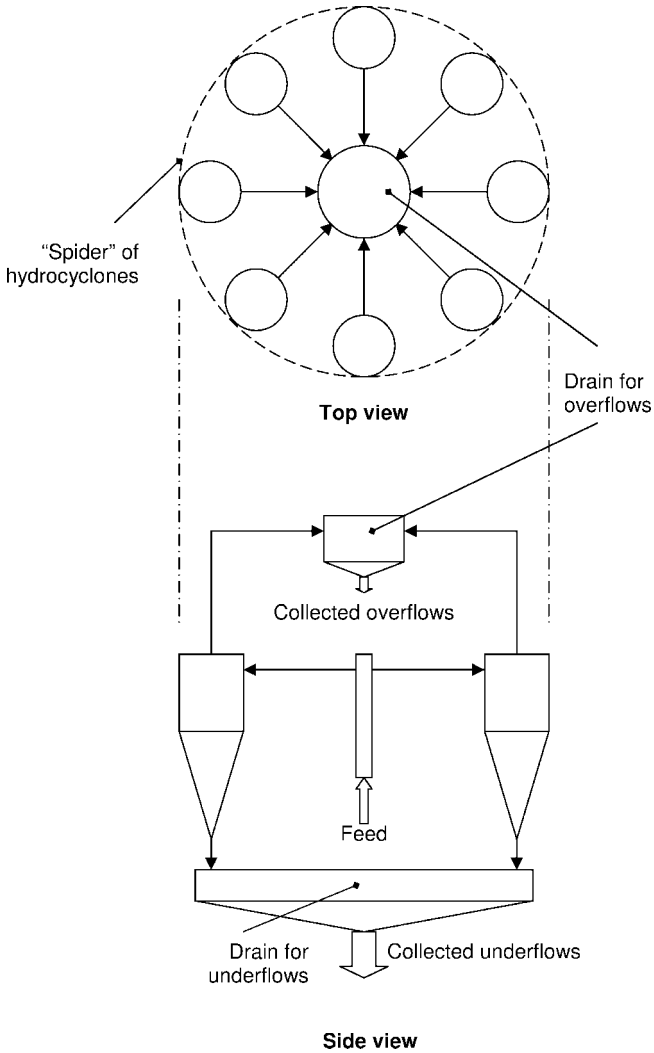

Figure 11. Diagram of radial (spider) arrangement of hydrocyclones for industrial scale

clone bodies. For cyclone cavities with a cylindrical diameter of 10 mm the diameter of the deck is 300 mm. The whole cyclone deck can be molded by injection in one piece. The feed zone is the annular space closest to the periphery. The overflow is through peripheral screw-in-vortex finders. The underflow leaves via the central circular space where the small apex openings enter.

The design feed flow per cyclone cavity is about 250 L/h corresponding to a pressure drop of 600 kPa, approximately. A deck with 20 cyclones in parallel consequently may have a capacity of 5 m^3/h. This is, however, not the maximum capacity. The cyclone cavities can be closed off individually by insertion of a blinding device in order to control the capacity. For feed flows above 5 m^3/h, the decks can be stacked on top of each other and held together by connecting rods through the inner annular space. There is a limit to the number of decks that can be stacked on top of each other. If the stack of decks becomes too high, the pressure drop over the feed zone will result in a lower inlet pressure to the cyclones furthest downstream and, consequently, these cyclones will give lower throughputs and lower efficiency. Some other arrangements have been tested with different degrees of success. Some of them have been reviewed in [1].

As has been discussed, hydrocyclone technology has numerous applications in a wide variety of industries. Hydrocyclones have been particularly used for long time in mineral processing, as well as in the oil and gas industry. In these industries, many materials can be considered inert powders, and thus the equations of the dimensionless scale-up model proposed by Medronho and Svarovsky [49] can be employed to carry out experimental tests, as suggested above. Hydrocyclones have been also successfully used in other industries, such as the food and biochemical industry. Since most materials in these industries may behave as non-Newtonian, the model suggested by Ortega-Rivas and Svarovsky [51] would be more suitable for modeling purposes of processes involving biological systems.

Some examples of application of the dimensionless scale-up model are given below.

Example 1. Predict the cut size obtained if 1 L/s of a suspension containing 5 vol % of a 2150 kg/m^3 dolomite slurry in water, is going to be separated in a 25.4 mm diameter hydrocyclone of Rietema's geometry. The available pressure is 2×10^5 Pa and the underflow orifice is 5 mm.

The relatively low concentration of suspension would allow use of the density and viscosity of water as suspending medium properties. Also, since no temperature has been given, a reasonable assumption for the water density is 1000 kg/m^3 and for the water viscosity 0.001 kg m^{-1} s^{-1}. Thus, bearing in mind to use the superficial velocity in the cyclone body, $[4Q/\pi(D_c)^2]$ as the characteristic one, the following results are obtained:

The characteristic velocity is, firstly, obtained, as defined above, i.e.:

$$v = \frac{4\,(0.001)\,m^3/s}{3.1416(0.0254)^2 m^2} = 1.97\,m/s \tag{95}$$

Then, substituting into Equation 59, the Euler number is calculated as:

$$\mathrm{Eu} = \frac{2\left(2\times10^5\right)\,\mathrm{kg/s^2{\cdot}m}}{(1000)\,\mathrm{kg/m^3}\,(1.97)\,\mathrm{m/s}} = 102.7 \tag{96}$$

As stated before, due to the low suspension concentration, the set of equations suggested for this case [49] can be considered, and thus, using Equation 91 the underflow-to-throughput ratio may be obtained as:

$$\mathrm{R_f} = 1218\left(\frac{0.005\mathrm{m}}{0.0254\mathrm{m}}\right)^{4.75}(102.7)^{-0.3} = 0.1346 \tag{97}$$

Equation 89 can then be employed to compute the value of the *Stk*$_{50}$*(r)* group, i.e.:

$$\mathrm{Stk_{50}\,(r)} = \frac{0.047[\ln\,(1/0.1346)]^{0.74}\exp\,[8.96\,(0.05)]}{102.7} = 0.0019 \tag{98}$$

Finally, transposing the cut size from the definition of the Stokes number given by Equation 72, the result is obtained as follows:

$$\mathrm{x}_{50}'^2 = \left(\frac{(0.0019)\,(18)\,(0.001)\,\mathrm{kg/m{\cdot}s}\,(0.0254)\,\mathrm{m}}{(2150-1000)\,\mathrm{kg/m^3}\,(1.97)\,\mathrm{m/s}}\right) = 1.55\times10^5\,\mathrm{m} \tag{99}$$

The cut size calculated is approximately equal to 15 μm.

Example 2. A slurry containing calcium carbonate at a concentration of 15 vol % solids, is going to be clarified using hydrocyclones of Rietema's proportions. The slurry has been characterized being non-Newtonian in behavior, with values of the parameters n' and K' of 1.39 and 2×10^{-5} kg m$^{-1} \cdot$ s$^{-(2-n)}$, respectively. The density of the slurry has been determined as 1250 kg/m^3. It is required to clarify 8 L/s of slurry, counting on a maximum pumping capacity of 315 kPa. Find out the optimum cyclone diameter and the number of units, if necessary, in order to obtain a cut size of 8 μm. Small scale experimentation has determined values of some dimensionless relationships as follows: $Stk_{50}Eu = 0.0625$, and $Eu = 720$. The calcium carbonate density has been determined as 2800 kg/m^3.

The value of the parameter n corresponds to a shear-thickening fluid, while the requirement of using Rietema's cyclones makes it possible to use the model suggested by Ortega-Rivas and Svarovsky [51]. The maximum pumping capacity would produce a pressure drop of the same value (315 000 N/m^2), because the leaving streams discharge at atmosphere conditions. Substituting the definition of superficial velocity in the cyclone body, as used in the previous example, into Equation 59, and considering the density of the slurry due to its high concentration, the following relation is obtained:

$$D_c = \sqrt[4]{\frac{(Q^2)\,(Eu)\,(\rho_m)}{(1.23)\,(\Delta P)}} \tag{100}$$

Substituting, also, the same characteristic velocity in Equation 88 and transposing for the reduced cut size:

$$x_{50}^{n'+1}\,(r) = \frac{(Stk_{50}^*)\,(18)\,(3)^{n'-1}\,(K')\,[(2n+1)\,/\,(3n)]^{n'}\,(D_c)^{5-2n'}}{(1.27Q)^{2-n'}\,(\rho_s-\rho)} \tag{101}$$

As previously stated, the cut size is directly proportional to the hydrocyclone diameter, while this diameter is inversely proportional to the flow rate. It would be, therefore, necessary to verify whether the required cut size could be obtained in a single hydrocyclone, whose diameter would match the capacity needed. If both variables do not coincide (cut size and flow rate with a single unit), iterations would be needed until a number of units is found, each of them separating the required size, while processing the appropriate fraction of the total capacity. A first approximation would consist in calculating the hydrocyclone diameter, using the first equation derived above, which will handle the 8 L/s of slurry to be treated. Once having that diameter, with the second equation, the cut size that the single unit separates, will be known. If such diameter is wider than 8 µm, the volumetric flow rate will be fractioned in equal parts to recalculate hydrocyclone diameter until a match of number of units and cut size is found. The first approximation leads to the following values:

$$D_c = \sqrt[4]{\frac{(0.008)^2\,m^6/s^2\,(720)\,(1250)\,kg/m^3}{(1.23)\,(315000)\,kg/s^2{\cdot}m}} = 0.11\ m \tag{102}$$

and:

$$x_{50}^* = \sqrt[2.39]{\frac{\left(8.68{\times}10^{-5}\right)(18)\,(1.53)\left(2{\times}10^{-5}\right)kg/m{\cdot}s^{2-n}\,(0.872)\,(0.007445)\,m^{5-2n}}{(0.0608)\,(m^3/s)^{2-n}\,(1800)\,kg/m^3}} \tag{103}$$

$$x_{50}^* = 15\,\mu m \tag{104}$$

So it is found that a 0.11 m diameter hydrocyclone will separate particles much coarser than the required cut size. By doing iterations, when fractioning the rate four times, i.e., when $Q = 0.002\,m^3/s$, a diameter of hydrocyclone $D_c = 0.055\,m$ and a cut size of 7.8×10^{-6} m are found. Therefore, the election would be the use of four units, 5.5 cm in diameter, to handle the capacity required and obtain the cut size needed.

6. References

1. L. Svarovsky: *Hydrocyclones*, Holt Saunders, Eastbourne, UK 1984.
2. D. Bradley: *The Hydrocyclone*, Pergamon Press, Oxford 1965.
3. J. Happel, H. Brenner: *Low Reynolds Number Hydrodynamics*, Prentice-Hall Inc., Englewood Cliffs., NJ, USA 1965.
4. A. Einstein, *Ann. Phys.* **19** (1906) 289.
5. H. L. Frisch, R. Simha: "The Viscosity of Colloidal Suspension and Macromolecular Solutions", in F. Eirich (ed.): *Rheology*, Vol. **1**, Academic Press, New York 1956.
6. E. Guth, R. Simha, *Kolloid Z.* **74** (1936) 226.
7. R. Simha, *J. Appl. Phys.* **23** (1952) 1020.
8. D. G. Thomas, *Ind. Eng. Chem.* **55** (1963) 18.
9. S. F. Lin, R. S. Brodkey,*J. Rheol.* **29** (1985) 147.
10. R. G. Griskey, D. G. Nechrebecki, P. J. Notheis, R. T. Balmer, *J. Rheol.* **29** (1985) 349.
11. A. J. De Vries, *Rheol. Abs.* **5** (1985) 32.
12. A. B. Metzner: "non-Newtonian Technology: Fluid Mechanics, Mixing, and Heat Transfer", in T. B. Drew, J. W. Hoopes Jr. (ed): *Advances in Chemical Engineering*, vol. **1**, Academic Press, New York 1959.
13. A. B. Metzner: "Flow of non-Newtonian Fluids", in V. Streeter (ed.): *Handbook of Fluid Mechanics*, McGraw-Hill Book Co., Inc, New York 1961.
14. D. W. Dodge, A. B. Metzner, *AIChE J.* **5** (1959) 189.
15. R. G. Shaver, E. W. Merril, *AIChE J.* **5** (1959) 181.
16. D. G. Thomas, *AIChE J.* **8** (1962) 373.
17. H. F. Trawinski, *Filtr. Sep.* **6** (1969) 7.
18. R. R. Irani, C. F. Callis: *Particle Size Measurement: Interpretation and Applications*, John Wiley & Sons, London 1963.
19. P. H. Fahlstrom, *Proc. Inst. Min. Metall.* **69** (1960) 632.
20. M. I. G. Bloor, D. B. Ingham, *Filt. Sep.* **21** (1984) 266.
21. H. Schubert, T. Neesse: "A Hydrocyclone Separation Model in Consideration of the Turbulent Multi-phase Flow", in BHRA Conference on Hydrocyclones Proceedings, Paper 3, Cambridge, UK 1980.

22. T. Allen: *Particle Size Measurement*, Chapman & Hall, London 1981.
23. G. V. Barbosa-Canovas, E. Ortega-Rivas, P. Juliano, H. Yan: *Food Powders: Physical Properties, Processing and Functionality*, Kluwer Academic/Plenum Publishers, New York 2005.
24. C. C. Harris, *Powder Technol.* **5** (1971/1972) 39.
25. H. J. E. v. Tenbergen, K. Rietema: "Efficiency of Phase Separations", in K. Rietema, G. C. Verver (eds.): *Cyclones in Industry*, Elsevier Publ. Co., Amsterdam 1961.
26. L. Svarovsky: "The Efficiency in Separation Processes", in R. J. Wakeman (ed.): *Progress in Filtration and Separation*, Elsevier Publ. Co., Amsterdam 1979.
27. H. F. Trawinski: "Hydrocyclones", in D. B. Purchas (ed.): *Solid/Liquid Separation Equipment Scale-up*, Uplands Press Ltd., Croydon, UK 1977.
28. H. F. Trawinski, *Aufbereitungs-Technik*. **17** (1976) 449.
29. M. A. Doheim, G. A. Ibraheim, A. A. Ahmed, *Int. J. Miner. Process.* **14** (1985) 149.
30. K. Rietema, *Chem. Eng. Sci.* **15** (1961) 298.
31. D. A. Dahlstrom, *Trans. Amer. Inst. Min. (Metall.) Engrs.* **184** (1949) 331.
32. N. Yoshioka, Y. Hotta, *Chem. Eng. Japan.* **19** (1955) 632.
33. A. L. De Gelder: "Model Tests with Hydrocyclones", in *Proceedings of IChemE Symposium on Scaling-up of Chemical Plants and Processes*, London, 1957.
34. A. J. Lynch, T. C. Rao, *Indian J. Technol*. **6** (1968) 106.
35. H. F. Trawinski, *Chem. Ing. Tech*. **30** (1958) 85.
36. L. Svarovsky, B. S. Marasinghe: "Performance of Hydrocyclones at High Feed Solid Concentrations", in BHRA Conference on Hydrocyclones Proceedings, Paper 10, Cambridge, UK 1980.
37. T. Neesse, W. Dalman, D. Espig: "Effect of Turbulence on the Efficiency of Separation in Hydrocyclones at High Feed Solids Concentration", in BHRA Conference on Hydrocyclones Proceedings, Paper B3, Bath, UK 1984.
38. T. J. Nappier-Munn: "Influence of Medium Viscosity on the Density Separation of Minerals in Cyclones" in BHRA Conference on Hydrocyclones Proceedings, Paper 6, Cambridge, UK 1980.
39. R. D. Williamson, T. R. Bott, I. H. Newson: "Small Particle Separation in a Hydrocyclone", in BHRA Conference on Hydrocyclones Proceedings, Bath, UK 1984.
40. F. J. Fontein, J. G. Van Kooy, H. A. Leniger, *Brit. Chem. Eng.* **7** (1962) 410.
41. D. N. Moir, *The Chemical Engineer* **523** (1985) 20.
42. A. M. Gerrard, C. J. Liddle, *The Chemical Engineer* **407** (1975) 295.
43. A. Zanker, *Chem. Eng.* **82** (1977) 122.
44. T. C. Rao, K. Nageswararao, *Chem Eng.* **82** (1975) 121.
45. R. W. Day, C. N. Grichar: "Hydrocyclone Separation", in P. A. Schweitzer (ed.): *Handbook of Separation Techniques for Chemical Engineers*, McGraw-Hill Book Co, New York 1979.
46. S. J. Kline: *Dimensional Analysis as a Research Tool*, McGraw-Hill Book Co, New York 1962.
47. M. L. Wasserman, J. C. Slattery, *AIChE J.* **10** (1964) 383.
48. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena*, John Wiley & Sons, New York 2001.
49. R. A. Medronho, L. Svarovsky: "Tests to Verify Hydrocyclone Scale-up Procedure", in BHRA Conference on Hydrocyclones Proceedings, Paper A1, Bath, UK 1984.
50. M. Antunes, R. A. Medronho: "Bradley Hydrocyclones: Design and Performance Analysis", in L. Svarovsky, M. T. Thew (eds.): *Hydrocyclones: Analysis and Applications*, Kluwer Academic Publishers, Dordrecht, Netherlands 1992.
51. E. Ortega-Rivas, L. Svarovsky, *Fluid/Part. Sep. J.* **6** (1993) 104.

Metallic Foams

Markus Weber, Fraunhofer-Institute for Manufacturing and Advanced Materials (IFAM), Department of Casting and Foaming Technologies, Bremen, Federal Republic of Germany

John Banhart, Fraunhofer-Institute for Manufacturing and Advanced Materials (IFAM), Department of Casting and Foaming Technologies, Bremen, Federal Republic of Germany

Joachim Baumeister, Fraunhofer-Institute for Manufacturing and Advanced Materials (IFAM), Department of Casting and Foaming Technologies, Bremen, Federal Republic of Germany

1. Introduction 1
2. Production 1
2.1. Production from Metallic Melts 1
2.2. Production from Metal Powders . . . 3
2.3. Production by Deposition Techniques 6
3. Toxicology and Occupational Health . 8
4. Applications and Outlook 8
5. References 8

1. Introduction

Cellular materials are widespread in everyday life and are used for cushioning, insulating, damping, construction, filtration, and many other applications. Highly porous materials combine high stiffness with very low density. Cellular materials frequently occur in nature as construction materials (e.g., wood, bone). The fact that even metals and metal alloys can be produced as cellular solids or metal foams is not as well known as the possibility to foam more traditional engineering materials such as plastics, ceramics, and glass. Metallic foams offer interesting properties due to the combination of metallic character with a porous structure.

Since ca.1960, many attempts have been made to foam metals or to produce porous metallic structures. Here some of the more important and successful procedures are described.

There is no clear-cut and generally accepted definition for the term "foam". First, one has to distinguish between liquid and solid foams. A liquid foam is a fine dispersion of gas bubbles in a liquid. Cooling a liquid foam to below the melting point of the respective material yields a frozen liquid foam which is then clearly a solid foam. What is generally meant in the context of metallic foams are in general solid metallic foams. The use of the word "solid foam" could be restricted to materials which originally were in the liquid state. However, customarily other porous structures such as sintered metal powders are often also called foams, although they never were in a liquid state. Hence, the use of the word "foam" is often extended to porous metal structures which are not actually foams, but resemble foams with respect to their high porosity, the interconnectivity of the solid material, and their irregular structure.

2. Production

2.1. Production from Metallic Melts

One group of foam-making processes starts from the molten metal, which is processed into a porous material by foaming it directly, by using an indirect method via a polymer foam, or by casting the liquid metal around solid filler materials, which reserve space for the pores or which remain in the foam.

Direct Foaming of Melts. Metallic melts can be foamed directly under certain circumstances by injecting gases into the liquid. Normally, the gas bubbles which are then formed in the metallic melt will tend to rise to its surface quickly due to the high buoyancy forces in the high-density liquid, but this can be impeded by increasing the viscosity of the molten metal. This can be done by adding fine ceramic powders or alloying elements that form particles in the melt.

Numerous attempts to foam liquid metals were undertaken in the 1960s and 1970s [1], but apparently the processes then invented could not

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.c16_c01

be sufficiently optimized to yield foams of a satisfactory quality and cost. In the 1990s, however, a number of new developments have taken place, and better production routes are now available.

Currently, there are two methods for directly foaming metallic melts. One of these is exploited by Cymat (originally by Alcan) in Canada and Hydro Aluminum in Norway [2–9] for foaming aluminum and aluminum alloys. Silicon carbide, aluminum oxide, or magnesium oxide particles are used to enhance the viscosity of the melt. The first step involves making an aluminum melt and adding 5 – 15 % of one of these particulate substances. The main problems here resemble those encountered in the production of conventional metal-matrix composites (MMCs), namely, the wetting of the particles by the melt and achieving a homogeneous distribution of the reinforcing particles [10]. The liquid MMC melt is then foamed by injection of gases (air, nitrogen, argon) with specially designed rotating impellers that produce very fine gas bubbles and distribute them homogeneously in the melt. The resulting foam floats to the surface of the melt where it can be pulled off, for example, by means of a conveyor belt. Care must be taken not to damage the foam structure by excessive shearing of the semisolid foam. The resulting material can, in principle, be as long as desired, as wide as the vessel containing the melt, and typically 10 cm thick.

The foamed material is used in the as-cast state with a closed outer surface or is cut into the required shape after foaming. Due to the high content of ceramic particles, machining of MMC foams could be problematic. The advantage of the process is the ability to produce large volumes at relatively low cost, and the low density that can be achieved. Porosities range from 80 to 97 %. A possible disadvantage is the eventual necessity for cutting the foam, thereby opening the cells, and brittleness due to reinforcing particles in the cell walls. Attempts to make shaped parts by casting the semiliquid foam into molds or by shaping the emerging foam with rolls have been undertaken [11], [12]. Some properties of such metal foams have been investigated [13], [14]. MMC foams are probably the least expensive metallic foams.

The second method of foaming melts directly is to add a foaming agent to the melt instead [15], [16], [17]. The foaming agent decomposes under the influence of heat and releases the foaming gas. In the small-scale commercial process operated at Shinko Wire in Japan, calcium is used to increase the viscosity of the melt, presumably by forming an Al_4Ca intermetallic phase, and titanium hydride TiH_2, added by using a rotating impeller, serves as the foaming agent, releasing hydrogen gas on heating. The foamed aluminum is produced in batches of approximately $2000 \times 1000 \times 600$ mm by foaming and solidification in a closed mold. Typical densities are $0.25 - 0.33$ g/cm^3. The blocks are cut into sheets of the required thickness [18] (e.g., 5 – 250 mm, standard thickness 10 mm. The pore structure of the material, known as ALPORAS, is rather uniform [19]. Mechanical [19] and sound absorption properties [20] have been measured. The material is reportedly more expensive than the MMC foams.

Solid – Gas Eutectic Solidification (Gasars). A method developed in the Ukraine [21], [22] exploits the fact that some liquid metals form a eutectic system with hydrogen gas. Dissolving hydrogen in these metals under high pressure (typically 50 bar) gives a pore-free melt in which the hydrogen is completely dissolved. Reducing the temperature and pressure causes the melt to go through the two-phase region. Below the eutectic temperature, a two-phase field corresponding to solid plus gas is reached. If the process parameters (cooling rate and pressure profile) are chosen appropriately, the gas accumulates as fine gas bubbles in the solid, thus forming a foam. The possibility of solidifying the liquid directionally offers the advantage of making foams with elongated pores. If the casting vessel is cylindrical, radial and axial pores can be made [23]. The maximum porosities that can be achieved by this process are not very high (5 – 75 %), but metals with moderate to high melting points such as copper and nickel can be foamed. The pore structure of such foams, known as gasars, is somewhat problematic [24], [25], so that further improvements have to be awaited. The mechanical properties (compression and tension) of gasars have been determined [26], [27], [28].

Investment Casting. Metal foams can also be fabricated without directly foaming the metal [29–31] by starting from a polymer foam. The

polymer foam is turned into a structure with open pores by manipulating the foaming process or by a reticulation (thermophysical shock treatment to remove residual membranes). The foam is then filled with a slurry of heat-resistant material, for example, a mixture of mullite, phenolic resin, and calcium carbonate [30]. After drying, the polymer is removed (e.g., by heating) and molten metal is cast into the resulting open voids which exactly reproduce the original foam structure. After removal of the mold material (e.g., by high-pressure water), a metallic foam is obtained that is an exact image of the original polymer foam. Figure 1 shows a micrograph of such a foam, which is being sold by ERG in California under the trade name DUOCEL. Various grades are available ranging from 2.5 to 16 pores per centimeter. Complex shaped parts can be fabricated by preforming the polymer foam. Prices are reportedly high. Aluminum alloys are usually used, but other metals can also be processed. The densities and foam morphologies are determined by the polymer foam. Porosities typically range from 80 to 97 %.

Syntactic Foams with Filler Materials. Lightweight porous metals can be produced by casting around inorganic granules or hollow spheres of low density or by infiltrating such materials with a liquid melt. A loose bulk of expanded clay granules, foamed glass spheres, or hollow aluminum oxide spheres can be used [32], [33]. The granules are then introduced into the melt, or the melt is poured over the bulk filler material. The heat capacity and conductivity of the granules is very low and therefore does not disturb the flow of the metal too much. Due to the high surface tension of the liquid metal, wetting of the granules is a problem, and in general the interstices between the granules will not be filled completely. Applying a slight vacuum or external pressure facilitates infiltration significantly. After solidification of the melt, the filler can optionally be removed, for example, by heating. A wide range of metals can be processed in this way, including aluminum, magnesium, zinc, lead, and tin. Parts of a predefined shape can be fabricated by designing a mold of the appropriate geometry. Sandwich panels have also been made [34].

2.2. Production from Metal Powders

Instead of molten metal, metal powders can be used to make porous metallic structures. Several methods can be applied. In some of these processes the powders are processed into a compact precursor material prior to the actual foaming step, in others powders (or fibers, hollow spheres, etc., which are regarded as powders here) are used for direct processing.

Fraunhofer Process. Foamed metals can be produced by a powder metallurgical method invented and patented [35–39] by the Fraunhofer-Institute IFAM in Bremen, Germany. The process begins with the mixing of metal powders (elemental metals, alloys, or powder blends) with a foaming agent, after which the blend is compacted to yield a dense, semifinished product. In principle, compaction can be performed by any technique that ensures that the foaming agent is embedded in the metal matrix without any residual open porosity, for example, uniaxial compression, extrusion, and powder rolling [34]. The choice of compaction method depends on the required shape of the precursor material. However, extrusion seems to be the most economical method at the moment and is therefore the preferred technique. Rectangular profiles with various cross sections are usually made. Thin sheets are then obtained by rolling.

Heat treatment at temperatures near the melting point of the matrix material is the next step. During this process the foaming agent, which is homogeneously distributed within the dense metallic matrix, decomposes. The released gas forces the compacted material to expand, forming a highly porous structure. Prior to foaming, the precursor material can be processed into sheets, rods, profiles, and other shapes by conventional techniques like rolling, swaging, or extrusion in order to improve the flow conditions during foaming in molds. The density of the metal foams can be controlled by adjusting the content of foaming agent and several other foaming parameters such as temperature and heating rate. If metal hydrides are used as foaming agents, a content of less than 1 % is sufficient in most cases.

Although most applications are based on aluminum foams, the Fraunhofer method is not restricted to this metal: tin, zinc, brass, lead,

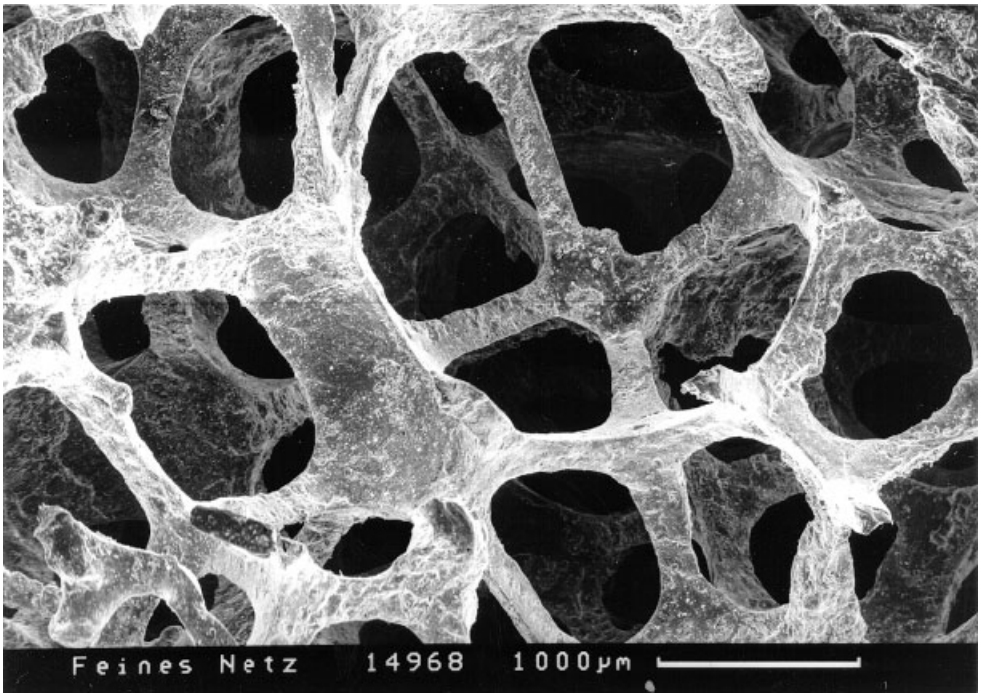

Figure 1. SEM image of a metal foam made by investment casting (DUOCEL)

and some other metals and alloys can also be foamed by choosing appropriate foaming agents and process parameters. The most common alloys for foaming, however, are pure aluminum, 2xxx alloys and 6xxx alloys. Casting alloys such as AlSi7 and AlSi12 are also frequently used because of their low melting point and good foaming properties. In principle virtually any aluminum alloy can be foamed by properly adjusting the process parameters. Figure 2 shows a typical cross section of a powder metallurgical foam. The random distribution of cell sizes and shapes is typical for a foaming process.

Foaming a piece of precursor material in a furnace results in a lump of metal foam with an undefined shape unless expansion is restricted in certain directions. This is done by inserting the precursor material into a hollow mold and expanding it by heating. In this way near-net-shape parts can be prepared. Quite complicated parts can be manufactured by injecting the expanding foam into suitable molds and allowing the final expansion to occur there [40].

Sandwich panels consisting of a foamed metal core and face sheets can be obtained by gluing the face sheets to a sheet of foam. Alternatively, if a pure metallic bonding is required, conventional metal sheets (aluminum or steel) are roll-cladded onto a sheet of foamable precursor material [39]. The resulting composite can be deformed in an optional step, for example, by deep drawing. The final heat treatment then leads to sandwich structures such as those shown in Figures 3 and 4. The large aluminum/aluminum foam part shown in Figure 4 was a joint development of the car component supplier Karmann and the Fraunhofer-Institute for a concept car in which structural aluminum foam applications were demonstrated. The part is intended to replace the rear seat wall and has higher stiffness and lower weight than the conventional steel part.

Tubes can be filled with aluminum foam in various ways (Fig. 5). The easiest way is to insert a rod of foamable precursor material into the tube that is to be filled and to place the tube into a furnace. The precursor material foams and

Figure 2. Lead foam made from metal powder

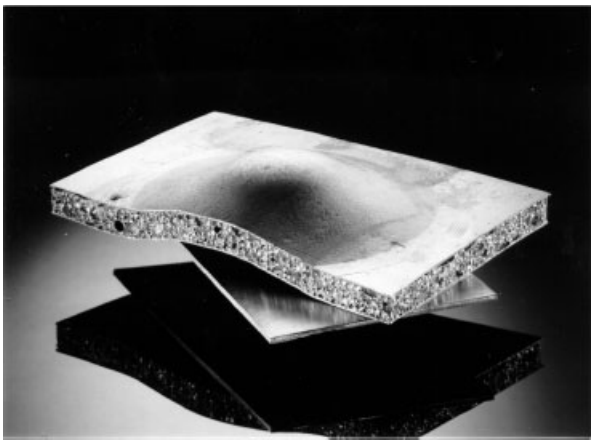

Figure 3. Composite structure made of an aluminum foam core and steel face sheets (dimensions $250 \times 250 \times 20$ mm)

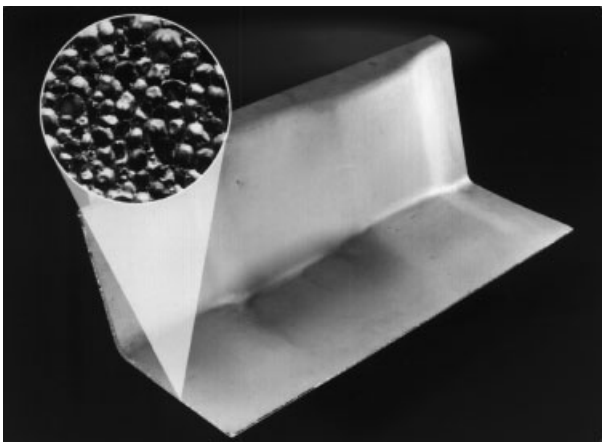

Figure 4. Shaped sandwich of aluminum foam with aluminum face sheets (width 1000 mm, thickness 8 mm)

eventually fills the tube completely. The disadvantage of this method is that only tubes with a significantly higher melting temperature than the material to be foamed can be used. If aluminum tubes are filled with foam in this way, the top of the tube melts during foaming because it is not supported mechanically and the heat from the furnace is not conducted into the precursor material. An alternative method is to insert a foamable tube into the tube that is to be filled. Then the foam expands towards the center of the tube. The foamable material is always in contact with the outer tube and therefore supports it. The expansion of the foam can be limited to a certain wall thickness by using a core in the center of the tube. The third possibility involves a co-extruded structure consisting of an outer tube of conventional aluminum and an inner tube of foamable precursor material. The two layers have a metallic bond, in contrast to the situation in which the inner tube is merely inserted into the outer one. This bond ensures better heat flow between the two tubes and therefore minimizes the risk of melting the outer tube. Moreover, the metallic bond between the foam and the outer tube is also better, because oxide layers between the inserted and outer tubes partially prevent a metallic bonding during foaming. Examples of foam-filled tubes are shown in Fig 6.

More details of the Fraunhofer process are given elsewhere [41–43]. Many mechanical and other physical properties have been determined [44–46]. The process is now being extended to metals with higher melting points such as iron and steel [47–49]. Due to their excellent biocompatibility, titanium foams could be used in prosthetic applications. However, these are still in the developmental stage.

Gas Entrapment. Metals can be foamed without a propellant by compressing powders into a precursor material and allowing gas to be entrapped in the metal structure during compaction [50], [51]. Heating the precursor material then leads to an expansion of the metal due to the internal pressure created by the entrapped gas. The process has mainly been used for making porous titanium structures. Titanium powder is filled into a can that is then evacuated and refilled with argon gas. The filled can is densified by hot isostatic pressing, subsequently worked, and finally foamed by heating. Refinements and extensions of this process have been developed, and properties of materials and components have been determined [52], [53].

Foaming of Slurries. Metallic foams can be produced by preparing a slurry of metal powder mixed with a foaming agent. The slurry is poured into a mold after mixing and dried at elevated temperature. The slurry becomes more viscous and starts to foam as gas begins to evolve. If sufficient stabilizing measures are taken, the expanded slurry can be dried completely to give a metal foam. Such foams have been produced from aluminum powders with orthophosphoric acid/aluminum hydroxide or hydrochloric acid as foaming agents [54]. Relative densities as low as 7 % have been achieved, but there are problems with low strength [55] and cracks in the foamed material.

Other Techniques. There are many ways to make porous metallic products from metal powders, fibers, or hollow spheres. The easiest method is to sinter loose powder fillings in a canister to give a porous material with open porosity. Mixtures of metal powders and polymer binders can be extruded and then heat treated to produce porous parts [56]. In this way porous materials with cylindrical pores can be produced. Reaction sintering of metal powder mixtures can also yield porous products.

Hollow spheres made of titanium or steel can be used to form highly porous structures by sintering [57]. Ordered and disordered arrangements can be realized. The density of such structures is extremely low with strengths that are still sufficient for many applications. By infiltrating the interstices between the hollow spheres, the strength can be increased.

2.3. Production by Deposition Techniques

Deposition techniques use metal ions as starting materials. The metal is galvanically deposited on a polymer foam with open cells [58–60]. Galvanic deposition on a polymer foam requires a certain electrical conductivity of the initial polymer foam. This is achieved by dipping the polymer foam into graphite solutions or by coating

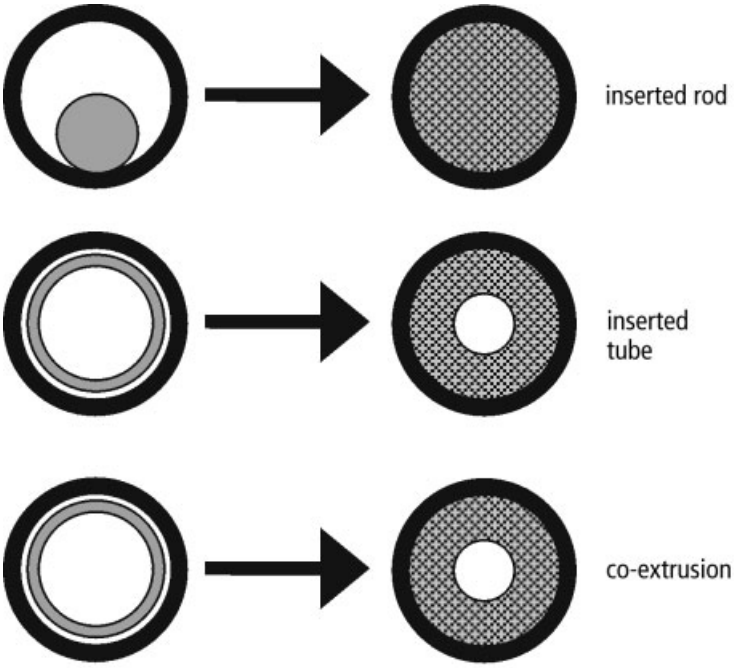

Figure 5. Methods for filling metal tubes with metal foam [39]

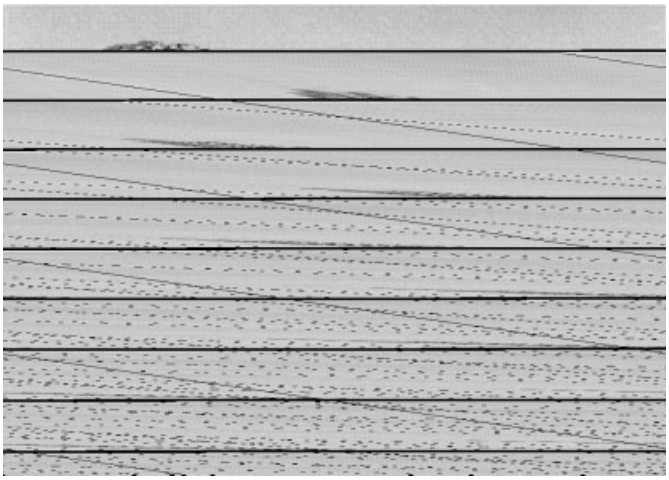

Figure 6. Foam-filled aluminum tubes (Foam: AlSi12 alloy, tube: 6061 alloy)

it with a thin conductive layer by metal vaporization. After electroplating, the polymer can be removed from the metal – polymer composite by thermal treatment.

Foams of various grades can be fabricated, ranging from 2 to 30 cells per centimeter. The preferred metal is nickel or an NiCr alloy, but copper foams can also be fabricated. Foams have been offered on a commercial basis under the name RETIMET (Dunlop, Great Britan) and CELMET (Sumitomo, Japan [56]). There is also a source in Belorussia [60]. Sheets with thick-

nesses between 2 and 10 mm are available with densities ranging from 0.2 to 0.5 g/cm^3.

3. Toxicology and Occupational Health

Typically, metal foams have the same toxicological properties as the corresponding dense, base metal alloy. No specific precautions for the processing and handling of metallic foams are known.

4. Applications and Outlook

The development of aluminum-based foams has been especially promoted by the requirement for certain structural lightweight applications so that the widest range of foams is available for this metal and its alloys. Various small-scale production facilities for aluminum foams are now existent in Japan, Canada, Slovakia, Austria and Germany. Some prototypes of aluminum foam parts have been developed and tested. However, it is not yet entirely clear in all cases whether the aluminum foam part can compete with the conventional part that it is intended to replace both with respect to its physical and mechanical properties and to its economical performance.

An interesting application for three-dimensional shaped sandwich foam structures was recently presented by the car outfitter Karmann. These complex shaped panels will be used as fire walls or luggage-compartment walls and will simultaneously stiffen the entire automotive body structures [61]. A comparison of a luggage compartment wall made of conventional stamped steel with an equivalent aluminum foam sandwich panel reveals a ten times higher stiffness of the latter. A combination of this high-stiffness material with the aluminum space frame concept opens the opportunity to build different variations of vehicles – even cabriolets with their high stiffness requirements – based on one common body structure.

The cost effectiveness of the various process is difficult to evaluate and should be regarded on a part by part basis. Only few methods are suitable for high-volume mass production, and many will remain restricted to specialized applications where cost is not the most important parameter. However, some of the processes for the production of aluminum foams, especially the melt and the powder metallurgical processes, have the potential for mass production at reasonable costs. There seems to be a market for these materials in applications where the unique properties of the metallic foam are fully exploited. This is the case when the foam has multiple functions, for example, when its low weight in combination with its high-temperature resistance is desired.

5. References

1. US 3 794 481, 1974 (L. M. Niebylski, C. P. Jarema, P. A. Immethun).
2. US 4 973 358, 1990 (I. Jin, L. D. Kenny, H. Sang).
3. WO-A 91/03578, 1991 (I. Jin, L. D. Kenny, H. Sang).
4. WO-A 92/03582, 1992 (I. Jin, L. D. Kenny, H. Sang).
5. US 5 112 696, 1992 (I. Jin, L. D. Kenny, H. Sang).
6. WO-A 94/172218, 1994 (M. Thomas, L. D. Kenny).
7. J. T. Wood, *Proc. Fraunhofer USA Metal Foam Symposium, Oct. 7. – 8. 1997, Stanton, Del.,* 31.
8. WO-A 90/00115, 1990 (W. Ruch, B. Kirkevag).
9. WO 91/01387, 1991 (W. Ruch, B. Kirkevag).
10. WO-A 91/19823, 1991 (D. J. Lloyd, A. D. McLeod, P. L. Morris, I. Jin).
11. WO-A 94/09931, 1994 (L. D. Kenny, M. Thomas).
12. WO-A 92/21457, 1992 (H. Sang, L. D. Kenny, I. Jin).
13. O. Prakash, H. Sang, J. D. Embury, *Mat. Sci. Eng. A* **199** (1995) 195.
14. Product data sheets of Hydro Aluminium, Sunndalsøra 1994.
15. EP-A 0 210 803 A 1, 1986 (S. Akiyama et al.).
16. US 4 713 277, 1987 (S. Akiyama et al.).
17. M. Itoh, T. Miyoshi, *Mat. Res. Soc. Symp. Proc.* **521** (1998) R3.2.
18. ALPORAS product information, Shinko Wire Co. Ltd., Izumisano 1998.
19. M. Otsuka, A. Kojima, M. Itoh, E. Ishii, in "Science and Engineering of Light Metals", *Proc. Conf. RASELM '91, Tokyo, Oct. 1991* (1991) 999.

20. E. Ishii, M. Itoh, Y. Morisawa, *Kobe Steel Engineering Reports* **41** (1991) 59.
21. US 5 181 549, 1991 (V. I. Shapovalov).
22. MRS-Bulletin, April 1994, p. 24 – 28.
23. Y. Zheng, S. Sridhar, K. C. Russel, in S. Komareni et al. (eds.): "Advances in Porous Materials", *MRS Society Bull.* **371** (1995) 365.
24. A. Pattnaik, S. C. Sanday, C. L. Vold, H. I. Aaronson, in "Advances in Porous Materials", *MRS Society Bull.* **371** (1995) 371.
25. J. M. Wolla, V. Provenzano, in S. Komareni et al. (eds.): "Advances in Porous Materials", *MRS Society Bull.* **371** (1995) 377.
26. V. Provenzano, J. Wolla, P. Matic, A. Geltmacher, A. Kee, in S. Komareni et al. (eds.): "Advances in Porous Materials", *MRS Society Bull.* **371** (1995) 383.
27. A. E. Simone, L. J. Gibson, *Acta. Met.* **44** (1996) 1437.
28. *J. Mat. Sci.* **32** (1997) 451.
29. DUOCELL product information, ERG Inc., Oakland Ca., 1996.
30. Y. Yosida, C. Hayashi, *Proc. Conf. Casting Sci. Technol.* 1990, Sept., 103.
31. I. Wagner et al., DGM Hauptversammlung 20. – 23.5.1997, Braunschweig, Germany, (abstract).
32. DE 1 933 321, 1971 (W. Thiele).
33. W. Thiele, *Metals Mat.* 1972, Aug., 349.
34. M. Hartmann, R. F. Singer, in *Proc. Symp. Metal. Foams, Bremen March 6. – 7. 1997,* 39.
35. DE 4 018 360, 1990 (J. Baumeister).
36. US 5 151 246, 1992 (J. Baumeister).
37. EP 0 460 392 A 1, 1996 (J. Baumeister).
38. DE 4 101 630, 1991 (J. Baumeister, H. Schrader).
39. DE 4 426 627, 1997 (J. Baumeister, J. Banhart, M. Weber).
40. F. Simancik, F. Schörghuber, *Mar. Res. Soc. Symp. Proc.* **521** (1998) R3.7.
41. M. Weber, J. Baumeister, J. Banhart: "Metals and Composites," *5th Eur. Conf. Adv. Mat.* (1997) 467.
42. J. Banhart, J. Baumeister, M. Weber, *Proc. Euro. Conf. Adv. PM Mat. (PM '95), Birmingham Oct. 23. – 25.,* 1995, 201.
43. F. Simancik, J. Kovácik, F. Schörghuber, Proc. cited in [34], p. 171.
44. J. Banhart, J. Baumeister, *J. Mat. Sci.* **33** (1998) 1431.
45. H. P. Degischer, U. Galovski, R. Gradinger, R. Kretz, F. Simancik, Proc. cited in [34], p. 79.
46. M. Weber, Phd thesis 1995, TU Clausthal ISBN 3-9805748-3-0.
47. B. Kriszt, A. Falahati, H. P. Degischer, Proc. cited in [34], p. 59.
48. M. Weber, M. Knüwer, Proc. cited in [34], p. 73.
49. C. J. Yu, H. Eifert, M. Knüwer, M. Weber, *Mat. Res. Soc. Symp. Proc.* **521** (1998) R3.4.
50. R. L. Martin, R. J. Lederich, *Metal Powder Report* 1992, Oct., 30.
51. M. W. Kearns, P. A. Blenkinshop, A. C. Barber, T. W. Fathing, *Int. J. Powder Met.* **24** (1988) 59.
52. D. S. Shih et al., *Mat. Res. Soc. Symp. Proc.* **521** (1998) R7.3.
53. D. S. Shih et al., *Mat. Res. Soc. Symp. Proc.* **521** (1998).
54. J. P. Drolet, *Int. J. Powder Met.* **13** (1977) 223.
55. S. B. Kulkarni, P. Ramakrishnan, *Int. J. Powder Met.* **9** (1973) 41.
56. L. Tuchinskiy, R. Loutfy, Proc. cited in [7], p. 23.
57. K. M. Hurysz et al., *Mat. Res. Soc. Symp. Proc.* **521** (1998) R5.3.
58. Product data sheet of SEAC International B.V., Krimpen aan den IJssel, The Netherlands 1986.
59. Product data sheet of Sumitomo Electric, 1986.
60. Product data sheet of Ecosot, Belorussia 1996.
61. H.-J. Emmelmann, W. Bunsmann, W. Seeliger, J. Baumeister, Proc. IBEC'97, Stuttgart Sept. 30. – Oct. 02 1997.

Nanostructured Materials

Bernard H. Kear, Department of Ceramic and Materials Engineering Rutgers – The State University of New Jersey, Piscataway, NJ 08855, United States

Ganesh Skandan, Nanopowder Enterprises Inc., Piscataway, NJ 08854-3908, United States

1. Introduction 1
2. Characteristics of Nanostructured Materials 1
3. Production of Nanostructured Powders 2
3.1. Spray Conversion Processing 3
3.2. Sol – Gel Processing 3
3.3. Aqueous Solution Processing 3
3.4. Physical Vapor Synthesis 4
3.5. Chemical Vapor Condensation 5
3.6. Chemical Vapor Reaction 6
3.7. Liquid Flame Spray 6
4. Processing of Nanostructured Bulk Structures and Coatings 7
4.1. Equal Channel Angular Pressing . . . 7
4.2. Liquid Phase Sintering of *n*-WC/Co . 7
4.3. Thermal Spraying of *n*-WC/Co 8
4.4. High-Pressure Consolidation 8
5. Processing of Novel Nanostructured Materials 8
5.1. Composite Nanoparticles 8
5.2. Films with Nanoscale Porosity 8
5.3. High Green Density Compaction Method 9
6. Applications of Nanostructured Materials 9
7. Future Perspective 11
8. References 11

1. Introduction

This article deals with the synthesis and processing of nanostructured ceramics and ceramic – metal composites. Optoelectronic materials such as porous silicon, multilayer nanostructured thin films (→ Thin Films), biomimetic materials such as hydroxyapetite, and carbon fullerenes (→ Carbon, Chap. 7.) are not covered.

Materials with fine-scale structures have long been recognized to exhibit remarkable and technologically attractive properties. Interest has been growing in a new class of ultrafine-grained materials, called nanocrystalline, nanophase, or nanostructured materials (*n*-materials) [1–3], wherein the particle or grain size is 1 – 100 nm. This interest stems not only from the outstanding properties that can be obtained from such materials, but also from the realization that early skepticism about the ability to produce high-quality nanoscale powders at a competitive cost was unfounded. Today, dozens of methods are available for producing nanopowders from metals, ceramics, and cermets, some of which have already been commercialized. Accordingly, the emphasis is shifting from powder synthesis to powder processing, i.e., the challenge of making useful coatings or bulk structures from nanopowders. Nanostructured bulk materials and protective coatings present major opportunities for advances in materials properties and performance in a broad range of engineering applications.

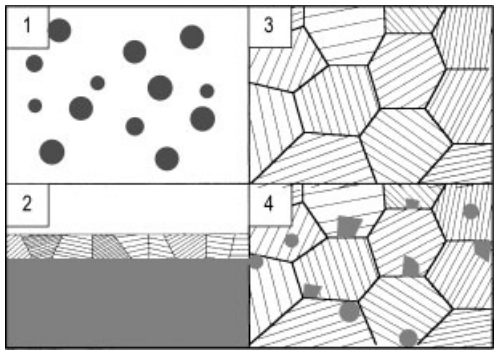

Figure 1. Schematic showing different classes of nanostructured materials (adapted from [4]).
1) Powder particles; 2) Nanostructured coatings/films; 3) Nanostructured monolithic bulk; 4) Nanocomposite bulk
1) – 4) can be porous or dense

2. Characteristics of Nanostructured Materials

As shown in Figure 1, the general term nanostructured refers to powders, coatings/films, and

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.c17_c01

bulk materials in which at least one of the phases is nanoscale [4]. In some cases, the individual powder particles are single crystals, while in others each particle may be composed of several nanocrystals. Nanocomposite materials include different combinations of metals, ceramics, and polymers. Furthermore, coatings and bulk structures can contain nanoscale closed or open porosity.

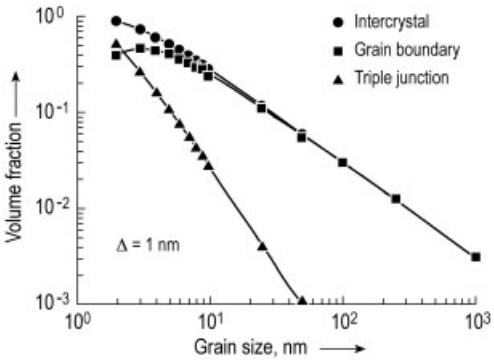

Figure 2. Fraction of atoms residing at grain boundaries as a function of grain size (width of grain boundary = 1 nm)

Table 1. Nanomaterial properties and corresponding grain/particle size effects

Nanomaterial property	Corresponding grain/particle size effects
Enhanced sinterability of metal and oxide ceramic powders	An ultrafine particle size (< 100 nm) gives rise to a high driving force for densification
Improved UV scattering	UV scattering is optimum at a particle size of 30 – 50 nm
Enhanced gas sensitivity	The gas sensitivity of a material is proportional to the surface area and hence becomes significant in nanomaterials
Very high hardness and wear resistance	Sintered nanophase metals and cermets display very high hardness and wear resistance because of their unique flow characteristics
Stable colloidal suspensions	Nanoparticles that are not agglomerated lead to increased stability of colloidal dispersions because of a smaller particle size in the liquid medium

A feature of nanoscale materials is the high fraction of atoms that reside at particle surfaces or grain boundaries. Figure 2 shows a plot of volume fraction of atoms that reside at the grain boundaries of a nanoscale material with decreasing grain size [5]. At a grain size of less than 20 nm, as much as 30 % of the atoms can be present at the grain boundaries. Accordingly, the interfaces play a crucial role in determining the mechanical, optical, and electrical properties of the material.

In general, nanoscale materials exhibit dramatic changes in properties, such as enhanced sinterability at low temperatures [6], [7], improved UV scattering [8], very high hardness and wear resistance [9], [10], enhanced gas sensitivity, [11], smaller particle size in colloidal suspensions, superior magnetic and dielectric strength, and enhanced optoelectronic properties [12]. Table 1 lists some of the unique characteristics of nanomaterials and the corresponding grain/particle size effects.

3. Production of Nanostructured Powders

Ceramic and ceramic – metal nanopowder synthesis techniques fall into two broad categories: (1) precipitation from salt solutions, and (2) condensation from the vapor phase. Specific methods have evolved that differ only in the means by which thermal and/or mechanical energy is provided to the precursor species. These include laser ablation [13], microwave plasma synthesis [14], precipitation from organometallic or aqueous salt solution [15], spray pyrolysis [16], spray conversion processing [17], and gas condensation with an evaporative source [18] or a chemical precursor [19]. At issue in each of these techniques are particle size, distribution, morphology, purity, and degree of agglomeration. While some of these processes, such as precipitation from a solution, produce very small individual particles (< 10 nm) that form larger agglomerates, others, such as spray conversion processing, produce micron-size particles consisting of grains in the 20 – 50 nm size range.

Described below are several processes in each category of nanopowder synthesis: (1) spray conversion processing (SCP) for WC/Co [17], sol – gel processing of multicomponent oxides, and aqueous solution processing (ASP) for K-doped MnO_2 [20], which involve precipitation from salt solutions, and (2) physical vapor synthesis (PVS) and chemical vapor condensation (CVC) for Al_2O_3, TiO_2, and Fe_2O_3, chemical vapor reaction (CVR) for nonoxide ceramics and liquid flame spray (LFS) for single-phase and multicomponent oxides, which involve con-

densation of nanoparticles from the gas phase. The PVS process involves evaporation of elemental species, such as Al and Fe, that are subsequently oxidized in the gas phase, while the CVC and CVR processes involve pyrolysis of vapors of organometallic compounds and halides, respectively, in a hot zone, followed by condensation. The LFS process involves injection of liquid precursor droplets into a combustion flame.

3.1. Spray Conversion Processing

Tungsten carbide/cobalt (WC/Co) hardmetal is composed of a high volume fraction (> 60 vol %) of hard WC particles that are cemented together with a much softer Co-rich binder phase. Typically, such cemented carbides are produced by mechanical mixing of the constituent WC and Co powders, followed by cold pressing and liquid-phase sintering. After sintering, the material has a characteristic bicontinuous structure, in which the WC and Co phases form interpenetrating networks in three dimensions. Such a composite structure displays high hardness, superior wear resistance, and good fracture resistance. Hence, WC/Co hardmetal is the material of choice for a host of applications, but particularly machine tools, drill bits, and wear-resistant parts.

The difficulty of uniformly mixing ultrafine WC and Co powders by mechanical means has heretofore limited the scale of the WC grain size attainable in the final sintered product to about 0.3 µm. Rutgers University and Nanodyne Inc. have developed an alternative powder processing technology, called spray conversion processing (SCP), which produces premixed powders on the submicron scale.

The production of *n*-WC/Co powders by SCP technology involves essentially three steps: (1) mixing of aqueous solutions of tungsten and cobalt salts to fix the composition of the starting solution, (2) spray drying of this solution to form a homogeneous precursor powder, and (3) fluid-bed thermochemical conversion (pyrolysis, reduction, and carburization) of the precursor powder to the desired *n*-WC/Co product powder. Using a pilot-scale production unit (Fig. 3), Nanodyne produces *n*-WC/Co powders with compositions extending over the range of commercial interest from 3 – 25 wt % Co.

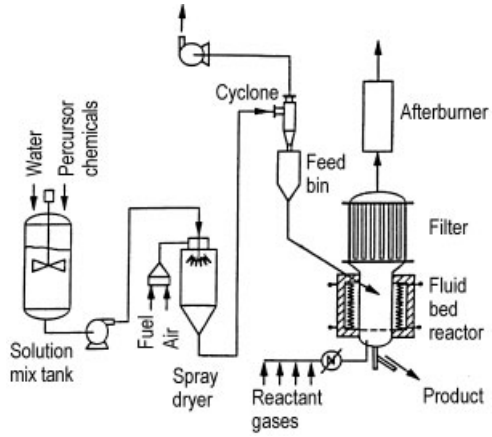

Figure 3. Schematic of Nanodyne's pilot-scale unit for the production of nanostructured WC/Co powder.
a) Solution mix tank; b) Spray dryer; c) Cyclone; d) Feed bin; e) Fluid bed reactor; f) Filter; g) Afterburner

3.2. Sol – Gel Processing

Sol – gel processing is a versatile method for producing not only multicomponent oxide nanopowders, but also for synthesizing nanocomposite particles. Bursill et al. [21] synthesized $BaTiO_3$ and $PbTiO_3$ nanoparticles (ca. 25 nm grain size) by the sol – gel method. Below a critical size, there is no paraelectric/ferroelectric phase transition in $BaTiO_3$. Estournes et al. [22] found that Ni nanoparticles dispersed in silica gel exhibit superparamagnetic behavior. Similarly, Piccaluga et al. synthesized $Fe_2O_3 - SiO_2$ nanocomposite particles by a gelation method using tetraethoxysilane and iron(III) nitrate as starting materials. Amorphous iron(III) oxide particles with a diameter of 3 – 4 nm were observed, which displayed superparamagnetic behavior.

To produce powders with a very high surface area (e.g., 700 m^2/g for SiO_2), Beaucage et al. [23] developed an aerosol – gel method in which the sol – gel chemistry is carried out in the gas phase.

3.3. Aqueous Solution Processing

Many methods exist for synthesizing nanopowders from aqueous solution precursors. For instance, metal chloride solutions can be reduced with sodium trialkylborohydride to form

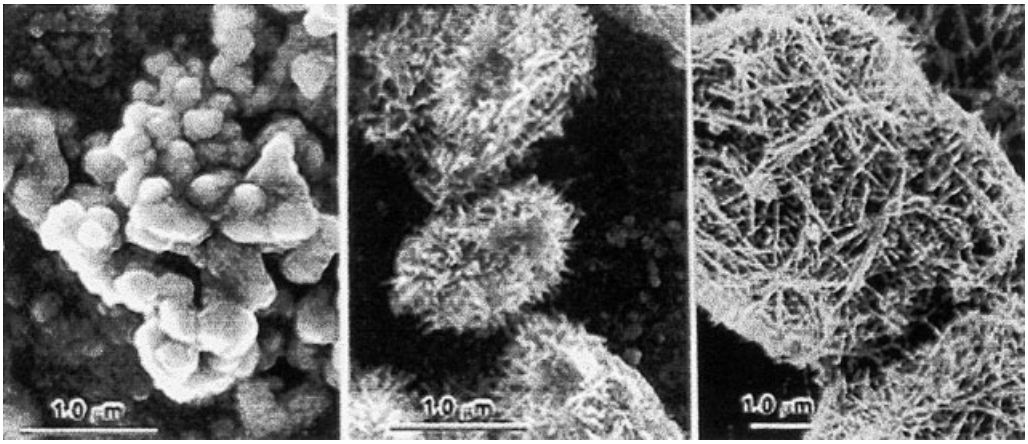

Figure 4. Micrographs of solution synthesized K-doped MnO_2 showing (a) initial nanoparticle agglomerated structure, (b) partially transformed structure after 4 h of reflux, and (c) completely transformed nanofibrous structure after 24 h of refluxing.

2 – 5 nm metal powders, e.g., *n*-Ni and *n*-Fe. Mixed metal chloride solutions can also be employed for synthesizing alloy powders, e.g., *n*-Ni/Cr and *n*-M50 steel [24] and hydroxide and oxide powders, e.g., *n*-Ni$(OH)_x$ and yttrium-stabilized zirconia (*n*-YSZ). In addition, a procedure has been developed for synthesizing large quantities of nanopowders by ultrasonically spraying an aqueous solution of mixed salts into a tank containing the reductive solution. This method permits solution dispersal of nanoparticles in a non-agglomerated form and continuous removal of reaction byproducts.

High surface area K-doped MnO_2 nanofibers can be synthesized by reacting aqueous solutions of $KMnO_4$ and $MnSO_4$ in the presence of nitric acid, by using a refluxing technique in combination with controlled solution mixing [20]. The initial nanoparticle agglomerates are gradually transformed into a nanofibrous structure (Fig. 4). Depending critically on the synthesis conditions, the fiber diameters range from 5 to 50 nm, with lengths up to a few micrometers and spacings in the range 20 – 200 nm. The K-doped MnO_2 nanofibers are *c*-axis oriented monocrystals, and have the cryptomelane-type hollandite structure with the approximate composition KMn_8O_{16}. In this structure, about half of the available K^+ sites are occupied and the rest are empty. For this reason, this type of manganese oxide is a molecular sieve that is particularly attractive for catalytic and ion-exchange purposes. In general, such intrinsically nanoporous materials make excellent cation-exchange materials. Moreover, with surface areas of > 250 m^2/g, they are promising catalysts for oxidation reactions and ionic/electronic conductors for rechargeable batteries.

3.4. Physical Vapor Synthesis

The synthesis of nanoparticles by the physical vapor synthesis (PVS) method, first introduced by Granqvist and Buhrmann [25], and Gleiter et al. [3], [26] (Fig. 5), involves evaporation of metal species under a reduced pressure of inert gas in a ultrahigh-vacuum chamber. Nanoparticles develop in a thermalizing zone just above the evaporative source due to interactions between the hot vapor species and the much colder inert gas atoms in the chamber. These particles are convectively transported and collected on a liquid-nitrogen-cooled stainless steel drum. Oxide ceramic powders are synthesized by oxidizing the metal nanoparticles in the gas phase or after deposition on the cold finger. Subsequently, Nanophase Technologies [18] developed a continuous process that eliminates the dependence on natural convection. In this process (Fig. 6), precursor material is introduced at a controlled rate into a chamber. In the chamber, a plasma arc is formed between a nonconsumable electrode and the precursor. The precursor, typically a high-purity metal rod, passes through the plasma arc and is melted and vaporized. A reactive gas, typically oxygen, is introduced into the chamber

and reacts with the evaporated precursor, causing nanoparticles of oxides to be formed upon condensation. Additional gas is turbulently introduced to accelerate cooling of the particles. The gas propels the nanoparticles into a collector housing which contains a filter that allows the gas to exit but traps the flaky agglomerated nanocrystalline particles. The nanoparticles that collect on the filter media are periodically harvested. Nanopowders of Fe_2O_3, Al_2O_3, TiO_2, and CeO_2 are routinely produced in tonnage quantities by the PVS method under the trade name Nanotek.

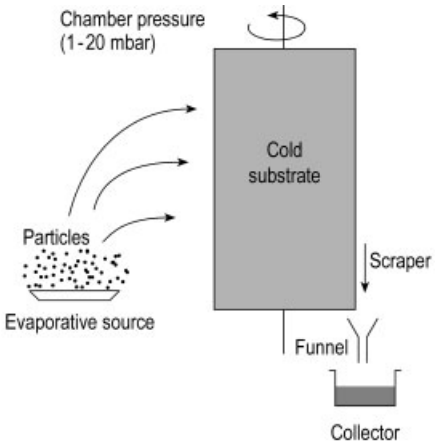

Figure 5. The original gas condensation process for producing nanoparticles, introduced by GLEITER et al. [3]

A less economical method of generating metal vapors is by laser evaporation of metal targets. Alternatively, oxide targets can also be ablated to form oxide nanoparticles directly. Several researchers have used Nd:YAG, Excimer, or a CO_2 laser to generate plumes of metal vapors, which are subsequently oxidized downstream to form nanoparticles [27–29].

3.5. Chemical Vapor Condensation

The synthesis of nanoparticles by the chemical vapor condensation (CVC) method, first introduced by CHANG et al. [19], involves controlled thermal decomposition of organometallic precursors in a reduced-pressure environment. By using a hot-wall reactor and an inert carrier gas for the precursor, nonoxide ceramic nanopowders can be synthesized. Using a combustion flame reactor allows oxide ceramics to be produced (Fig. 7). A unique feature of the combustion flame/CVC process is the specially designed burner which gives a completely uniform flat combustion zone with a uniform temperature profile and gas-phase residence time over its entire surface. Thus, nanoparticles are produced that are essentially monodispersed. Furthermore, by rapidly quenching the clusters/nanoparticles as they emerge from the flame, non-agglomerated powders can be generated.

Low-pressure flat flames had not previously been investigated for purposes of nanopowder production. One study performed at low pressure [30] yielded some important information but focused on very low precursor concentrations that minimally affect the chemical flame structure. In addition, this work was performed in an open flame without a substrate. Such flames are characterized by a significantly different thermal flame structure without a quench zone after the peak flame temperature.

In contrast, atmospheric flames have been studied at great length and are routinely used for powder synthesis. The low-pressure stagnation-point flames being scaled by Nanopowder Enterprises Inc. offer some significant advantages over atmospheric-pressure synthesis [31]. First, high concentrations of flame radicals can be maintained at lower temperatures. This implies that precursor decomposition temperatures can be lower, resulting in less sintering of colliding particles. This in turn, leads to reduced agglomeration. Second, the residence times of the synthesized particles in the low-pressure environment are shorter. Both volumetric flow rates per unit burner area and distance to the stagnation plane are similar in atmospheric flame synthesis reactors, but the pressure in the combustion flame/CVC process is about 35 times lower, leading to a reduction in residence time of nearly an order of magnitude. This implies that (1) the synthesized nanoparticles in combustion flame/CVC spend substantially less time in the hot zone of the flame, and (2) the nanoparticles suffer fewer potentially agglomerating particle – particle collisions, since the overall temperatures are lower. Third, the inherent uniformity of the stagnation flow environment, coupled with the increasing strength of the thermophoretic force relative to the drag force at

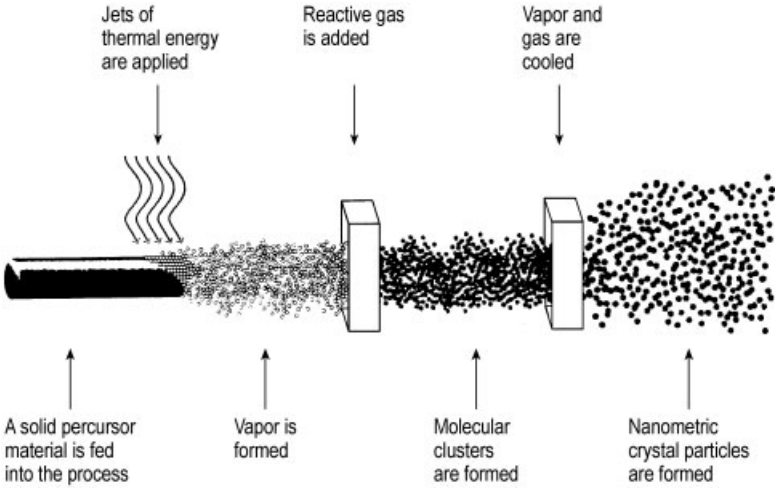

Figure 6. Schematic of the Nanophase Technologies Physical Vapor Synthesis (PVS) technique [18]

low pressures, suggests that particles which condense at different radial locations will experience similar trajectories and hence time/temperature histories.

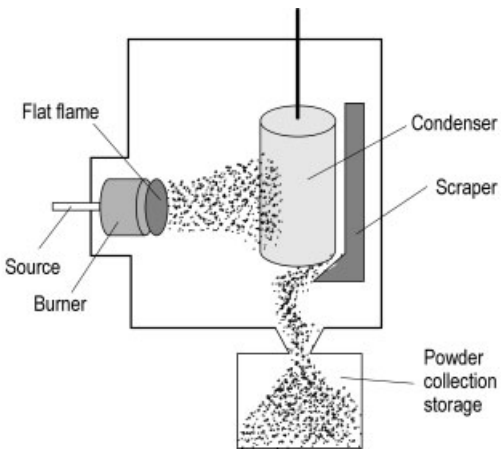

Figure 7. Schematic of the combustion flame/CVC process

3.6. Chemical Vapor Reaction

KONIG et al. [32] developed a chemical vapor reaction process at H. C. Starck and Co. for producing nanoparticles of metals and nonoxide ceramics. Precursor compounds, which are typically halides, are evaporated upstream in evaporators that are incorporated in the gas preheater. The precursor metal compounds and the reactants, which are typically H_2, CH_4, or NH_3, are introduced into the reactor through two coaxial nozzles in the form of coaxial laminar component streams. The laminar flow leads to a narrow distribution of residence times of the nuclei or particles. Thorough mixing between the flow streams is achieved by incorporating an obstacle in the otherwise strictly laminar flow, so as to generate a Karman vortex path. To prevent the reactants from being deposited on the reactor wall, to which there is a strong energetic bias, the reaction medium is screened from the reactor wall by introducing an inert gas layer. The reaction occurs at temperatures between 500 and 2000 °C, in accordance with the following examples:

$$2\,TiCl_4 + 2\,NH_3 + H_2 \longrightarrow 2\,TiN + 8\,HCl$$

$$TiCl_4 + CH_4 \longrightarrow TiC + 4\,HCl$$

$$2\,TaCl_5 + 5\,H_2 \longrightarrow 2\,Ta + 10\,HCl$$

3.7. Liquid Flame Spray

Recognizing the need to deliver precursors at high rates into a hot zone for pyrolysis, KARTHIKEYAN et al. [33] have developed a process, called liquid flame spray (LFS), in which a liquid precursor chemical is atomized, and the atomized liquid droplets are used as the spray

feedstock for producing nanoparticles. Two fluid atomizers were incorporated along the axis of an oxy-hydrogen flame spray gun. The atomizer produces micrometer-sized droplets of the liquid jet and injects them into the hot zone of the combustion flame. Nitrates, acetates, and organometallics can be dissolved in an appropriate solvent, such as isopropyl alcohol, and injected into the combustion flame. As the liquid droplets are accelerated by the flame, heat transfer from the flame to the droplets leads to evaporation of the solvent, condensation of the precursor, chemical reaction, and formation of particles in the gas phase. While the process allows the formation of a high-surface oxide material, control of particle size is an issue, as the atomizer produces a range of droplet sizes, and different droplets have different thermal histories in the spray flame, depending on their size and injection velocities.

4. Processing of Nanostructured Bulk Structures and Coatings

4.1. Equal Channel Angular Pressing

Efforts at producing dense materials from discrete metal nanopowders have met with limited success, primarily due to the inability to control oxygen contamination. Metallic nanoparticles have been produced in an ultrahigh-vacuum chamber and compacted in situ to prevent oxygen contamination [34–36]. While it has been possible to reduce the oxygen content to less than 0.5 atom % in materials such as copper, such a complicated process is less likely to be viable on an industrial scale.

Valiev et al. [37], [38] have developed an alternative approach to forming dense nanocrystalline metallic materials. Instead of consolidating powders by traditional powder metallurgy techniques, the new process, called equal channel angular (ECA) pressing, relies on severe plastic deformation to refine the microstructure. One can either start with an ingot or ball-milled powders. By extruding the material through intersecting channels of identical cross section, full consolidation and grain-size refinement are achieved simultaneously. Nanocrystalline materials of Pb – Sn, Zn – Al, and Al – Cu with a grain size of less than 100 nm, have been achieved by this process. Such materials are superplastic and display exceptional mechanical properties [39].

4.2. Liquid Phase Sintering of *n*-WC/Co

n-WC/Co powders synthesized by the SCP method are densified by liquid-phase sintering with the addition of a grain-growth inhibitor, such as VC or Cr_3C_2. In VC-doped *n*-WC/Co materials, the hardness increases with increasing VC concentration up to a maximum of 2190 VHN at 0.8 wt % VC (Fig. 8). These data correlate with a reduced mean free path for the cobalt binder phase, i.e., reduced WC grain size. Measurements also show that nanograined materials possess superior hardness at all compositions. The relatively high hardness (> 1850 VHN) at high Co content (> 10 wt %) is not accompanied by a decrease in fracture toughness [40].

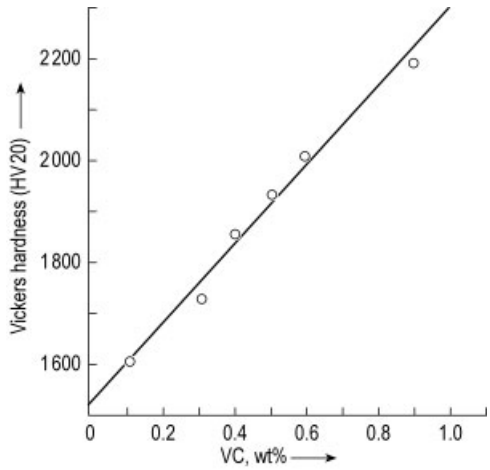

Figure 8. Hardness versus VC (grain growth inhibitor) content in WC/7 wt % Co alloys

Another interesting observation is the striking difference in response of micrograined and nanograined materials to a scratch test. The micrograined material shows evidence of combined plastic deformation and fracture of the WC grains, whereas the nanograined material yields by pure plastic deformation, despite its high hardness. Evidence for superior abrasive wear resistance of *n*-WC/Co, as well as for improved cutting performance of *n*-WC/Co drill bits has been obtained [21].

4.3. Thermal Spraying of *n*-WC/Co

Thermal spraying is a widely used industrial process for applying protective coatings to materials surfaces. The coating process involves feeding prealloyed powders into a flame or plasma gun (Fig. 9). The particles are rapidly heated to form a spray of partially or completed melted droplets. The large impact forces created as these particles arrive at the substrate surface promotes strong particle to substrate adhesion and the formation of a dense coating. An attractive feature of the process is its ability to produce coatings of almost any desired material, ranging in thickness from 25 µm to several millimeters, at relatively high rates.

Nanostructured coatings of hardmetals such as WC/Co offer significant potential improvements in abrasive wear resistance and hardness compared with conventional WC/Co coatings. Preliminary work on thermal spray deposition of Nanocarb WC/12Co and WC/15Co has yielded promising results.

4.4. High-Pressure Consolidation

Using a high-pressure/high-temperature (HPHT) hot pressing unit of advanced design, developed by VORONOV et al. [41], a new class of WC/Co/diamond nanocomposites are being produced [42]. The HPHT unit consists of shaped anvils of tungsten carbide and supporting prestressed steel rings. The anvils compress a shaped insert made of lithographic stone, which transmits a near-isostatic pressure to the reaction cell. The reaction cell consists of a resistively heated graphite crucible, which is capable of reaching 2000 °C in a few minutes.

During high pressure/low temperature (HPLT) sintering, the nucleation rate of the stable phase is increased, thereby accomplishing metastable/stable phase transformation at relatively low temperature [43], [44]. For example, at 5 GPa and 600 °C, nanoparticles of γ-Al_2O_3 transform completely into α-Al_2O_3. This transformation-assisted consolidation approach is being used to fabricate fully dense materials from oxide nanoparticles.

5. Processing of Novel Nanostructured Materials

Apart from the four major powder production technologies of commercial importance (see Chap. 3), several processing methods have developed for specific materials.

5.1. Composite Nanoparticles

VOLLATH et al. [45] have developed a two-stage microwave plasma technique in which nanoparticles formed near the inlet of the reactor are coated downstream with a layer of a second material, e.g., an Al_2O_3 layer on a ZrO_2 particle. The structure of these nanocoated particles depends strongly on the crystallization behavior of the phases forming the kernel and the coating. The main applications of these new nanocoated particles may either be seen in the formation of diffusion barriers to avoid grain growth or in the modification of physical properties of the core and the chemical properties of the surface.

CHE et al. [46] have developed a modified spray pyrolysis method for the synthesis of SiO_2-encapsulated Pd nanoparticles. These composite nanoparticles were formed from a palladium nitrate solution containing ultrafine SiO_2 particles by ultrasonic spray pyrolysis. A precursor particle formed below 700 °C in the drying stage was composed of a homogeneous mixture of *n*-SiO_2 and $PdO \cdot H_2O$. When PdO was decomposed above 700 °C, metallic Pd nanoparticles were formed in the SiO_2 matrix. Because of the high surface free energy, the metallic Pd particles coagulate and condense in the interior of the composite particle. As a result of the relocation within the composite particle, SiO_2 is forced out of the particle toward the surface, and an SiO_2-encapsulated Pd particle is formed. Similar structures of $Ag-SiO_2$ and $Cu-SiO_2$ were prepared by CRANE et al [47].

5.2. Films with Nanoscale Porosity

SiO_2 films with nanoscale porosity have attractive properties such as low dielectric constant

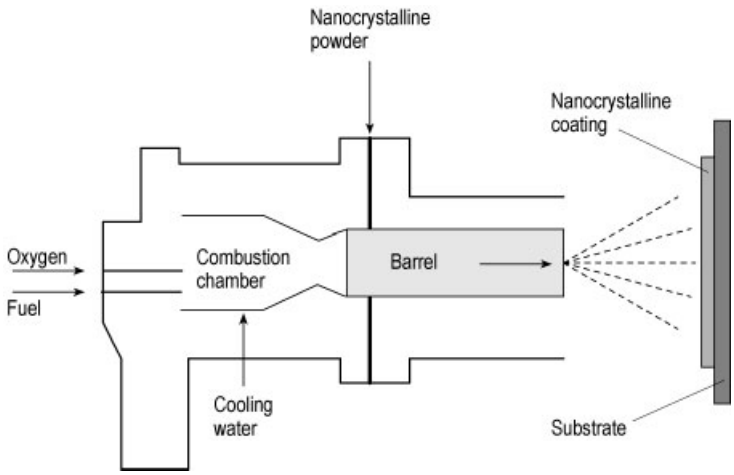

Figure 9. Schematic of the thermal spray coating process along with the various processing parameters.

($\kappa < 2$) and low density. Such films find applications as the interlayer dielectric material in devices. Jin et al. [48] have developed a sol – gel method to form such nanoporous thin films, in which processing parameters, such as process flow for deposition and post-deposition curing, relative rates of reaction, gelation, aging, and drying, are controlled to improve mechanical properties and dielectric constant.

Skandan et al. [49] have developed a low-pressure flame deposition (LPFD) process, which utilizes the flat-flame burner described in Section 3.4. LPFD is a one-step deposition process that completely avoids the use of powders as starting material. This is beneficial in terms of significant cost reduction, as well as the ability to easily access the critical thickness range of 1 – 10 µm. The LPFD process is capable of depositing films at rates greater than 1 µm/min, up to any thickness that one desires. The experimental arrangement for operating the flat flame in film-forming mode is similar to the powder-forming mode, except that the cold plate is replaced with a heated substrate. The "superheated" clusters or nanoparticles emerging from the hot zone of the flame are allowed to impinge upon a heated substrate. The nanoparticles sinter as they arrive at the substrate and form a film or coating. The desired extent of nanoscale porosity is obtained by controlling the number density of nanoparticles in the gas phase and the temperature of the substrate.

5.3. High Green Density Compaction Method

High green density is a prerequisite for achieving a high sintered density. Due to the increased surface area in nanopowders, it is not possible to achieve high green densities when the powder is pressed at conventional pressures. Ivanov et al. [50] have developed a pulsed densification method for oxides such as Al_2O_3 and ZrO_2, in which compression waves of 100 – 500 µs in duration and up to 2.5 GPa in amplitude are generated by a pulsed magnetic press. Green compacts with densities of up to 0.8 of the theoretical density have been reported. Similar results have been obtained by cold compaction under a pressure in excess of 5 GPa [51]. Higher packing densities may be attributed to interparticle sliding and plastic deformation.

6. Applications of Nanostructured Materials

Nanopowders are used (1) as dispersed phases in fluid media, (2) on supporting substrates, (3) for consolidation into dense structures, and (4) as feedstock for the formation of films and deposition of coatings.

Some of the notable areas where nanoparticles, nanostructured sintered materials, and

nanostructured films/coatings are being used or under development are as follows.

Chemical mechanical polishing slurries, containing nanoparticle dispersions of SiO_2, Al_2O_3, and CeO_2: The field of chemical mechanical polishing (CMP) of semiconductor materials is experiencing dramatic growth. While CMP was originally a process for expensive microprocessor production by major IC manufacturers that could support its cost, it is now used for an ever widening array of products, from modem chips to DRAMs. Manufacturers are beginning to use CMP in special niches, such as silicon on insulator wafer fabrication and multichip module assembly. While the equipment, materials and process technology form the different components of the CMP process, the slurry itself is a critical element. The particle composition, size and distribution, dispersability, morphology, shape, and hardness contribute to the polishing performance. According to the National Technology Roadmap Interconnect requirements projected until 2005, there is a need for reduction in particle size extending well into the nanoscale range (< 100 nm). This is because smaller size allows finer surface finish, while non-agglomeration and narrow size distribution assure that the slurry does not scratch the surface.

Ultraviolet scatterers such as TiO_2 and ZnO have maximum UV scattering efficiency in the particle size range 30 – 50 nm. Therefore, sunscreens containing nanoparticles have a correspondingly higher sun protection factors. Furthermore, provided they are non-agglomerating, the stability of the suspension is also improved.

Cosmetic pigments such as Fe_2O_3 in the form of nanoparticles can be used in cosmetic applications because of their finer texture and ability to spread easily on a surface.

Conducting inks containing nanoparticles of Cu and Ag enable a lower sintering temperature: Cu and Ag are among the few metallic nanoparticles that are not pyrophoric and can be handled in air. Such conducting ultrafine particles are dispersed in a liquid medium and used to form fine conducting lines on plastics and ceramic substrates. The low sintering temperature (about one-third of the melting point) allows low-temperature processing, especially when dealing with polymers.

Sintered nanostructured WC/Co: As-synthesized nanophase WC/Co is milled, compacted, and liquid-phase sintered to produce a variety of tools for mining and drilling operations. The high hardness and wear resistance increases the efficiency of the drilling operation.

IR windows: Sapphire (single-crystal Al_2O_3) is presently the material of choice for IR window applications because of its superior optical and mechanical properties compared to Y_2O_3, MgO, and AlON. However, processing single-crystal components is expensive. Polycrystalline Al_2O_3, AlON, and AlN with a nanocrystalline grain size of less than one-tenth of the wavelength of radiation is being considered as a potential replacement for sapphire. The nanomaterial is expected to have superior mechanical properties and comparable optical properties, compared to sapphire.

Sputtering targets: Sputtering is a widely accepted industrial process for depositing thin films. The quality of the sputtered film depends greatly on the properties of the target, including density, composition, and grain size. Using nanoparticles as starting materials is likely to yield very high density targets that have an ultrafine microstructure and a uniform composition.

Dielectric materials: For the past several years, much emphasis has been placed on low-temperature processing of thick-film ferroelectric materials for applications in multilayer ceramic capacitors (MLCCs), transducers, sensors and actuators for the following reasons: (1) fine-grained (0.7 – 1.0 μm) ferroelectric materials that have high permittivity and very good piezoelectric properties (e.g., $\varepsilon = 4000 - 6000$ at room temperature in $BaTiO_3$, and $d = 500$ pC/N in $Pb(ZrTi)O_3$) can only be synthesized by low-temperature processing, (2) when the fabrication temperature is less than 900 °C, reactions between the electrodes and the dielectric can be significantly reduced, and (3) low-temperature processing allows replacement of the expensive AgPd electrodes with cheaper Ni and/or Cu electrodes.

Gas sensors: Thin films of oxides such as SnO_2 and TiO_2 containing nanoscale porosity have applications as gas sensors. Thin film sensors for monitoring CO, CO_2, NO_x, SO_x, and other gases are usually fabricated by depositing dense materials that have a columnar structure. The current trend is to fabricate integrated microelectronics-based gas sensors as (1) it is more economical in the long run, (2) has a faster

response time, and (3) higher sensitivity. It has already been demonstrated that a dense film with a nanoscale grain size has a faster response time. It is expected that a porous film with the grains and pores in the nanoscale regime will have an even faster response time.

Low-voltage displays: One of the areas of development for the next generation of flat panel displays is phosphor powders that have high efficiency at a low operating voltage. Doped nanoparticles of oxidic materials such as Y_2O_3:Eu and Y_2O_2S:Eu have higher cathodoluminescence efficiency at operating voltages (ca. 200 V) than their coarse-grained counterparts. It is expected that further enhancements in the performance by way of surface passivation will lead to higher photoluminescence and electroluminescence efficiencies.

Engine components: Si_3N_4 reinforced with fine particles of SiC has evolved as a high-temperature structural material. Reducing the scale of Si_3N_4 and SiC has led to improved strength and toughness. In addition, the material displays superplastic behavior.

7. Future Perspective

Major advances have been made in the science of nanophase materials, and significant progress has been made on the technological front. In particular, impressive improvements have been made in the high rate processing of nanopowders at a competitive cost. Looking ahead, major innovations can be expected in the preparation of high-performance nanophase coatings and consolidated parts. The most immediate pay-off should be in the processing of thermally sprayed WC/Co-base coatings and liquid-phase sintered parts. On the other hand, in view of recent advances in high-pressure consolidation of nanoceramic powders, it is evident that within a few years, commercialization of sintered nanophase ceramics will be realized.

The use of nanopowder in dispersed form is an established technology in the catalysis industry. Further improvements in catalytic performance can be expected as a consequence of the ability now to produce nanoparticles that are tailored with respect to composition, structure, and morphology. In addition, dispersing nanoparticles in fluid media presents new opportunities that are being intensely pursued. A good example is the use of aqueous suspensions of monodispersed ceramic nanoparticles as abrasive medium for planarization of semiconductor and optical devices.

8. References

Specific References

1. R. Birringer, U. Herr, H. Gleiter, *Trans. of Jpn. Inst. Met.* **27** (1986) 43.
2. H. Hahn, J. Eastman, R. Siegel, *Suppl. Ceram. Trans.* **1** B, (1988) 1115.
3. H. Gleiter, *Progress in Mat. Sci.* **33** (1990) 4.
4. R. W. Seigel, *Nanostruct. Mater.* **3** (1993) no. 1, 1.
5. J. R. Weertman, *Scripta Metallurgica et Materialia* **24** (1990) 1351.
6. G. Skandan, H. Hahn, M. Roddy, W. R. Cannon, *J. Am. Ceram. Soc.* **77** 7, (1994) 1706.
7. H. Gleiter, *Nanostruct. Mater.* **1** (1992) no. 1, 1.
8. J. Robb, L. Simpson, *Drug Cosmet. Ind.* **3** (1994) 32.
9. L. E. McCandlish, B. H. Kear, B. K. Kim, *Nanostruct. Mater.* **1** (1992) no. 2, 119.
10. K. Jia, T. E. Fischer, *Wear* **200** (1996) 206.
11. N. Yamazoe, *Sensors and Actuators* **B5** (1991) 7.
12. G. S. Tompa et al., *MRS Symp. Proc.* **358** (1995) 701.
13. G. F. Gaertner, H. Lydtin, *Nanostruct. Mater.* **4** (1994) no. 5, 559.
14. D. Vollath, K. Sickafus, *Nanostruct. Mater.* **1** (1992) 427.
15. D. Gallaghar, W. Heady, J. Racy, R. Bhargava, *J. Mater. Res.* **10** (1994) 4, 870.
16. G. Messing, S. Zhang, V. Jayanthi, *J. Am. Ceram. Soc.* **76** (1993) 11, 2707.
17. B. H. Kear, L. E. McCandlish, *Nanostruct. Mater.* **3** (1993) 19.
18. Q. Ford, *Ceram. Ind.* **31** (1993).
19. W. Chang, G. Skandan, H. Hahn, S. C. Danforth, B. H. Kear, *Nanostruct. Mater.* **4** (1994) 345.
20. D. Reisner, T. D. Xiao, P. R. Strutt, Proc. of 13th Annual Battery Conference on Applications and Advances, Long Beach CA, Jan. 13 – 16, 369 (1998).
21. L. A. Bursill, B. Jiang, J. L. Peng, W. L. Zhong, P. L. Zhang, Technical Report, DE97-626946.

22. C. Estournes, T. Lutz, J. Happich, T. Quaranta, P. Wissler, J. L. Guille, *J. Magn. Magn. Mater.* **173** (1997) nos. 1 – 2, 83 – 92.
23. J. Hyeon-Lee, G. Beaucage, S. E. Pratsinis, *Chem. Mater.* **9** (1997) 2400.
24. K. E. Gonsalves, T. D. Xiao, G. M. Chow, C. C. Law, *Nanostruct. Mater.* **4** (1994) 130.
25. C. Granquist, P. Buhrmann, *J. Appl. Phys.* **47** (1976) 2200.
26. R. Birringer, H. Gleiter, H. P. Klein, P. Marquardt, *Physics Letters* **102A** (1984) No. 8, 365.
27. G. F. Gaertner, H. Lydtin, *Nanostruct. Mater.* **4** (1994) no. 5, 559.
28. R. Uyeda, M. Kato, *Technical Report of Toyota* **26** (1973) 66.
29. H. Lee, W. Riehmann, B. Mordike, *Z. Metallkunde* **84** (1993) 79.
30. D. Lindackers, M. G. D. Strecker, P. Roth, *Nanostruct. Mater.* **4** (1994) 545.
31. N. Glumac, Y-J. Chen, G. Skandan, *J. Mater. Res.* (in press).
32. US 5 356 120, (T. Konig, K. Bachle, H. Ewel, V. Rose, G. Zippenfenig, R. Klafki).
33. J. Karthikeyan, C. C. Berndt, J. Tikkanen, J. Y. Wang, A. H. King, H. Herman, *Nanostruct. Mater.* **8** (1997) no. 1, 61 – 74.
34. R. Birringer, U. Herr, H. Gleiter, *Transactions of the Japanese Institute of Metals* **27** (1986) Suppl. 43.
35. A. Tschope, R. Birringer, *Philosophical Magazine B* **68** (1993) 223.
36. P. Sanders, G. E. Fougere, L. J. Thompson, J. Eastman, J. R. Weertman, *Nanostruct. Mater.* **8** (1997) no. 3, 243 – 252.
37. A. V. Korznkov, I. M. Safarov, D. V. Laptionok, R. Z. Valiev, *Acta Metall. Mater.* **39** (1991) 3193.
38. R. Z. Valiev, R. S. Mishra, J. Groza, A. Mukherjee, *Scripta Mater.* **34** (1996) 1443.
39. R. S. Mishra, R. Z. Valiev, A. K. Mukherjee, *Nanostruct. Mater.* **9** (1997) 473 – 376.
40. K. Jia, T. E. Fischer, *Wear* **203 – 204** (1997) 310.
41. O. Voronov, A. Rahmanina, Proc. of IV Int. Symposium on Diamond Materials, Nevada, 1995.
42. R. Sadangi, O. Voronov, B. H. Kear, *CIMTEC '98*, 9th International Conference on Modern Materials & Technologies, Florence, Italy, June '98, in press.
43. S.-C. Liao, W. E. Mayo, K. D. Pae, *Acta Materialia* **45** (1997) no. 10, 4027.
44. S.-C. Liao, K. D. Pae, W. E. Mayo, *Nanostruct. Mater.* **8** (1997) no. 6, 645.
45. D. Vollath, D. Vinga Szabo, J. Hausselt, *J. Eur. Ceram. Soc.* **17** (1997) no. 11, 1317.
46. S. Che, D. Sakurai, T. Yasuda, K. Shinozaki, N. Mizutani, *J. Ceram. Soc. Jpn.* **105** (1997) no. 3, 269.
47. R. A. Crane, J. T. Matthew, R. P. Andres: "Nanophase and Nanocomposite Materials II," *Proc. Mat. Res. Symp.* (1997) 357.
48. C. Jin, J. D. Luttmer, D. Smith, T. Ramos, *Materials Research Society Bulletin* (1997) Oct., 39.
49. G. Skandan, N. Glumac, Y.-J. Chen, E. Heims, B. H. Kear, *J. Am. Ceram. Soc.* (1998) in press.
50. V. Ivanov, S. Paranin, A. Nozdrin, *Key Eng. Mater.* **132 – 136** (1997) 400.
51. S.-C. Liao, Y-J. Chen, B. H. Kear, W. E. Mayo, *Nanostruct. Mater.,* to be published.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

On-Line Monitoring of Chemical Reactions

Standard Article

Wolf-Dieter Hergeth[1]

[1]Wacker Polymer Systems, Burghausen, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.c18_c01.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Reaction Calorimetry
2.1. Introduction
2.2. Heat Flow Balance and Principles of Measurement
2.2.1. Heat Flow Balance
2.2.2. Basic Modes of Operation
2.3. Applications and Instrumentation
3. Ultrasonic Methods
3.1. Introduction
3.2. Theory
3.3. Applications
4. Dielectric Spectroscopy
4.1. Theory and Mechanisms
4.2. Instrumentation and Applications
5. Optical Spectroscopy
5.1. Introduction
5.2. Instrumentation for Reaction Monitoring
5.3. Applications of Optical Spectroscopy
5.3.1. UV – VIS Spectroscopy
5.3.2. NIR Spectroscopy
5.3.3. IR Spectroscopy
5.3.4. Raman Spectroscopy
5.3.5. Fluorescence
5.3.6. Other Spectroscopic Techniques
6. Particle Size Analysis
6.1. Scattering Techniques
6.1.1. Turbidimetry
6.1.2. Angular Static Light Scattering
6.1.3. Dynamic Light Scattering
6.1.4. Other Optical Techniques
6.2. Separation Techniques
7. Chromatography
7.1. Introduction
7.2. GC Hardware Components
7.3. Applications of Gas Chromatography
7.4. Other Chromatographic Techniques
8. Electroanalytical Methods
8.1. Introduction

8.2. Conductometry
8.3. Potentiometry
8.4. Amperometry, Voltammetry, and Coulometry
9. Miscellaneous Methods
9.1. Mass Spectrometry
9.2. Densimetry and Dilatometry
9.3. Rheometry
9.4. NMR Spectroscopy

[Top of Page]

1. Introduction

The terms “on-line monitoring” and “process analytical chemistry” are often used synonymously. They are gaining increasing attention and importance in both industry and academia. The extent of agreement between the two terms is indeed very broad in that they provide sufficiently accurate and immediate information on variables that describe the state of a chemical reaction.

On-line monitoring of reactions encompasses on-stream and on-reactor application of analytical methods to monitor the chemical composition of a reaction mixture, to identify process-related chemical species, and to quantify the concentration of reaction ingredients, products and byproducts. In addition to revealing the state of the reactor, on-line analysis of physical parameters (temperature, pressure, level, density, viscosity, etc.) may also reflect the extent of a chemical reaction.

Process analytical chemistry comprises applications which supply relevant process information of interest “in-time”: The time for sampling and analysis is very short compared to the overall reaction time and thus allows adequate monitoring and efficient control of the reaction. The utilization of the analytical data for control strategies makes process analytical chemistry an essential integral part of process engineering and control systems (see ⟶ Process Control Engineering; ⟶ Chemical Plant Design and Construction).

Process analytical methods may be classified as off-line, at-line, on-line, or in-line with respect to sampling, sample transport, and analysis itself (Table 1) [1]. There is no clear-cut line between the different classes, and the boundaries are even moving: “Some of today's off-line techniques may become tomorrow's on-line techniques” [2]. An increasing number of off-line techniques have been converted into on-line methods by automated, robot-assisted withdrawal of samples from the reactor or from bypass or process streams and feeding them into off-line instruments.

Table 1. Classes of process analyzers (adapted from [1])

Process analyzer	Sampling	Sample transport	Analysis
Off-line	manual	to remote or centralized laboratory	automated/manual
At-line	discontinuous/manual	to local analytical equipment	automated/manual quick check
On-line	automated	integrated	automated
In-line	integrated	no transport	automated
Noninvasive	no contact	no transport	automated

On-line monitoring and control of chemical reactions contributes to:

- Guaranteeing and improving product quality and consistency (i.e., repeatability of product properties within narrow specification ranges)
- Increasing the efficiency of the process
- Ensuring safe reactor operation by monitoring process and reactor parameters
- Understanding fundamentals of the reaction itself
- Saving time for analysis and sample transport
- Reducing emissions by avoiding sample withdrawal and transport
- Reducing costs for labor, raw materials, off-spec products, and process waste

In most cases, purchasing and installation costs of on-line and in-line analytical instruments exceed those of off-line equipment, but as soon as the measurements required exceed a certain (relatively small) number per day, on-line analysis becomes superior to off-line analysis (Fig. 1). However, the real savings of on-line reaction monitoring are due to improved process efficiency, lower raw materials consumption and waste generation, and, most important, the ability to manufacture high-quality products.

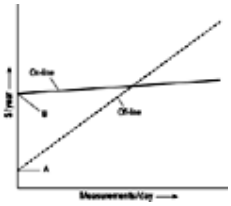

Figure 1. Total costs for on-line and off-line process analyses (including instrument costs, labor costs, energy, material, etc.)

A) Costs for purchase of off-line instruments; B) Costs for purchase and installation of on-line equipment

Several issues have to be addressed when designing an on-line analysis application [1], [3-8]:

- Information necessary to monitor and control the process (physical parameters, chemical composition, etc.)
- Frequency of measurements with respect to the timescale of the reaction
- Average values, typical fluctuations, dynamic ranges and expected extremes of properties
- Type, precision, and response time of sensors
- Robustness of sensors and simplicity of installation
- Full automation and minimum maintenance of equipment
- Number (combination) of sensors and location of measurement
- Proper sampling and, if required, sample conditioning
- Form of data output and further handling of information
- Compatibility with process control system
- Safety precautions and possible hazards
- Costs of instrumentation and availability of trained personnel

Demands on performance characteristics of on-line analytical techniques differ for academic research applications, process development and pilot plant operation, and monitoring of industrial manufacturing processes, as summarized in Table 2 (see also ⟶ Analytical Chemistry: Purpose and Procedures; ⟶ Chemometrics; ⟶ Plant and Process Safety; ⟶ Polymerization Processes; ⟶ Process Control Engineering; ⟶ Process Development; ⟶ Sampling).

Table 2. Relative importance of various performance characteristics of on-line analyses in academic research, process development, and industry [9]

Characteristics	Academic research	Process development	Manufacturing
High accuracy	+++	+++	+++
Good reproducibility	+++	+++	+++
Good selectivity	+++	+++	+++
Good sensitivity	+++	+++	+++
Extended linearity	++	+++	+++
Good stability	++	+++	+++
Robustness	++	+++	+++
High analysis frequency	++	+++	+/+++*
Short time delay of result	++	++	+++
Low price	+++	++	+
Multi-analyte analysis	+++	+++	+
Ease of use	+	+++	+++
Ease of validation	+	++	+++
High flexibility	+	+++	+
Ease of implementation	+	+++	++
Low maintenance	+	+++	+++

* Process-dependent. +++ Very important, ++ important, + not very important.

The number of textbooks introducing the field of process analytical chemistry is still limited [4], [6-8], [10-15]. Since 1993 a series of review articles on the progress in process analytics have been published [16-21]. On-line monitoring techniques play a major role in the contents of several scientific journals such as Journal of Process Analytical Chemistry [22], Applied Spectroscopy [23], Analytical Chemistry [24], and Analytical and Bioanalytical Chemistry [25]. Several international conferences devoted solely to process analytics have been established over the years. The most importants ones are IFPAC (International Forum for Process Analytical Chemistry [26]), APACT (Advances in Process Analytics and Control Technology [27]), and PREACH (Process Related Environmental Analytical Chemistry [28]). The reader should also be aware of the process analytical technology initiative of the FDA [29] which is of crucial importance to the pharmaceutical industry. Academic as well as academic-industrial consortia drive the development of new, more reliable and cost-effective techniques for on-line monitoring of chemical reactions applicable to the industry. The most active ones are CPAC (Center for Process Analytical Chemistry, Seattle, USA [30]), MCEC (Measurement and Control Engineering Center, Knoxville

USA [31]), CPACT (Center for Process Analytics and Control Technology, Newcastle, Strathclyde, Hull, UK [32]), MACC McMaster Advanced Control Consortium, Hamilton, Canada [33]), and CTAC (Control Theory and Applications Center, Coventry, UK [34]). A Working Group on Process Analytical Chemistry of the Society of German Chemists (GDCh, Frankfurt/M, D [35]) has been founded in 2005.

[Top of Page]

2. Reaction Calorimetry

2.1. Introduction

The majority of chemical processes are accompanied by temperature changes of the reaction mixture owing to release or consumption of heat in the course of the reaction. Heat evolution is a definite, reproducible, and directly measurable characteristic of a chemical reaction. This enables one to monitor the extent of a reaction by measuring temperature or heat flux changes (see [36-47] and ⟶ Plant and Process Safety for introduction, overviews, applications, and related subjects).

Reaction calorimetry is noninvasive, rapid, accurate, robust, and, as it is based on temperature measurements, relatively easy to carry out. Temperatures can be measured directly within the reaction mixture (reactor interior), at the reactor/exterior interface (reactor wall), or in the heating/cooling liquid of the reactor jacket.

An advantage of reaction calorimetry is that the heat of reaction Q_r derived from temperature measurements is directly proportional to the rate of reaction. This allows easy access to basic kinetic and thermodynamic data of chemical reactions.

With reaction calorimetry, rates of reaction or conversion of reactants can be determined quasi-instantaneously and continuously with a high degree of resolution. This makes reaction calorimetry an ideal tool for real-time feedback control of chemical composition during the course of reaction.

Reaction calorimetry not only provides quantitative information on the chemical process itself (e.g., heat of reaction, reaction rate, conversion) but also on reactor parameters necessary for safe reactor operation and process design:

- Global reaction kinetics
- Heat production rates
- Necessary cooling power
- Reactant accumulation
- Adiabatic temperature rise to avoid runaway reactions
- Heat-transfer coefficients for scale-up

However, reaction calorimetry is a nonselective method. It is impossible to distinguish between parallel chemical reactions with heat generation and simultaneous enthalpic processes within the system such as phase transitions, crystallization, mixing, and dissolution.

2.2. Heat Flow Balance and Principles of Measurement

2.2.1. *Heat Flow Balance*

The basis for reaction calorimetry is the energy balance around the reactor. Sources and sinks of heat Q_i which contribute to the overall heat flux balance (Eq. 1) are shown schematically in Figure 2:

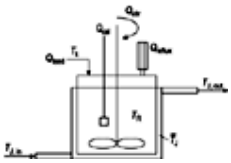

Figure 2. Schematic of a reaction calorimeter

$$\sum_i \dot{Q}_i = 0 \tag{1}$$

The quantity of interest in reaction calorimetry is the heat flux Q_r generated by chemical reactions within the reactor

$$\dot{Q}_r = V_R \sum_j r_j(-\Delta H_{r,j}) \tag{2}$$

where V_R is the volume of the reaction mixture, r_j the rate of reaction j, and $\Delta H_{r,j}$ the reaction enthalpy of reaction j. For typical heats of reaction, see ⟶ Plant and Process Safety.

Different rates of heat generation within the reactor and heat flow out of or into the reactor lead to temperature changes of the reaction mixture. This accumulated heat Q_{acc} is a major cause of uncertainty and error in quantitative analysis of reaction calorimetry data because it depends on the derivative of the reactor temperature T_R

$$\dot{Q}_{acc} = C_{p,tot}\dot{T}_R \tag{3}$$

where $C_{p,tot}$ is the heat capacity of the filled reactor (reactor plus reaction mixture).

Feeding of material into the reactor with a mass flow rate dm_k/dt (and also withdrawal of material from the reactor) leads to a heat flux dQ_{feed}/dt

$$\dot{Q}_{feed} = \sum_k \dot{m}_k c_{p,k}(T_k - T_R) \quad (4)$$

where $c_{p,k}$ is the specific heat capacity of the kth reactor feed and T_k its temperature.

The heat input by stirring the (viscous) reaction mixture Q_{stir} is given by

$$\dot{Q}_{stir} = P_0 N_v^3 d_{stir}^5 \varrho_{mix} \quad (5)$$

where P_0 is the power number, N_v the stirrer speed, d_{stir} the diameter of the stirrer, and ϱ_{mix} the density of the reaction mixture.

The energy balance of the reactor is also influenced by heat losses to the surroundings Q_{loss} (e.g., by conduction or radiation). It is important to identify and quantify all possible sources of heat losses around a reactor because heat losses might be significantly larger (2 – 5 times) than Q_{acc}, Q_{stir}, or Q_{feed}. In most cases, heat losses must be determined by separate experiments prior to the reaction or without a running reaction.

Additional heat can be removed by the insertion of reflux condensers into the reactor. The heat flow over the reflux condenser can be calculated according to Equation

$$Q_{reflux} = m_{cool} c_{p,cool}(T_{cool,in} - T_{cool,out}) \quad (6)$$

where $\dot{m}_{cool}$ is the coolant mass flow through the condenser, $c_{p,cool}$ is the specific heat capacity of the coolant, and $T_{cool,in} - T_{cool,out}$ the temperature difference across the condenser.

For calibration purposes (see below), heat may also be generated by additional heaters

$$\dot{Q}_{cal} = UI \quad (7)$$

where U and I are the voltage and current of the electrical heater, respectively.

The quantity that is actually measured in a calorimetric experiment is the heat flux through the reactor wall to the cooling/heating system Q_{trans}. There are two main methods to determine Q_{trans}: heat flux calorimetry and heat balance calorimetry.

Heat Flux Calorimetry. In heat flux calorimetry, the generated heat is measured by means of the temperature difference between reactor and jacket. The heat transfer between reactor interior and jacket medium Q_{trans} depends on the overall heat transfer coefficient k_Q through the wall, the heat exchange area of the reactor wall A_R that is actually in contact with the reaction mixture ("wet area"), and the difference between reactor temperature and mean jacket fluid temperature $\bar{T}_J$

$$\dot{Q}_{trans} = k_Q \cdot A_R(\bar{T}_J - T_R) \quad (8)$$

Typically, the temperature difference is sufficiently large for heat flux calorimetry to be very sensitive.

Advantages of heat flux calorimetry are its (i) high sensitivity because of the (relatively large) temperature difference between reaction medium and jacket fluid, (ii) fast response and high accuracy, and (iii) its independence on the flow rate of the jacket medium.

The reactor area for heat exchange with the jacket A_R is almost constant for batch reactions. There can be some variation in A_R as a result of density changes of the medium during the course of the reaction. In semibatch reactions, A_R is no longer a constant. Feeding of reaction components into the reactor leads to an increase of its filling level. If the feeding rates are known their influence on A_R can be taken into account.

A major problem in heat flux calorimetry is the determination of the exact k_Q value. The overall heat transfer coefficient may change considerably during the course of the reaction because it depends on, for example, the mixture's viscosity and density, the stirring rate (hydrodynamics of the reactor), and heat transfer through the reactor/jacket interface (influence of film formation or reactor fouling). Hence, $k_Q A_R$ must be calibrated separately in a typical heat flux calorimetry experiment. Several variations are proposed in the literature and offered by instrument manufacturers:

Continuous calibration during the reaction can be achieved by application of well-defined heat pulses produced by a calibration heater. The reactor heat measured with these pulses enables one to quantify the heat of reaction. However, the hot surface of the calibration heater may also cause problems with temperature-sensitive samples (e.g., product degradation, film formation).

REICHERT [48], [49] developed an elegant method to overcome the problem of variations in $k_Q A_R$, known as temperature oscillation calorimetry. A small sinusoidal temperature change is added to the overall temperature/time characteristic of the reactor via the jacket liquid or a calibration heater. These forced temperature oscillations

$$\dot{Q}_{\mathrm{osc}} \sim (1 + \sin(\omega t)) \tag{9}$$

create temperature oscillations of the reactor with amplitude δT_R and frequency $\omega = 2\pi/\tau$, where τ is the oscillation period. Decoupling and separate determination of oscillating and nonoscillating terms in the energy balance equation allows the simultaneous on-line calculation of the rate of reaction as well as the heat transfer value of the reactor $k_Q A_R$:

$$k_Q \cdot A_{\mathrm{R}} = -\omega C_{p,\mathrm{mixture}} \frac{\delta T_{\mathrm{R}} \sin \varphi_{\mathrm{r}}}{\delta T_{\mathrm{J}} - \delta T_{\mathrm{R}} \cos \varphi_{\mathrm{r}}} \tag{10}$$

where φ_r is the phase shift between the oscillations, and δT_J the amplitude of temperature oscillations of the jacket liquid. Temperature oscillation calorimetry has been critically reviewed and compared to conventional calorimetry in [50], [51].

One- or Two-Point Calibration. In the case of a constant heat transfer coefficient during the reaction the heat transfer value $k_Q A_R$ can be determined by calibration before the start of the reaction.

Several commercially available calorimeters apply a two-point calibration. The first calibration must be carried out before the reaction run starts to determine an initial value of $k_Q A_R$. A second calibration after completion of the reaction in combination with the assumption of a linear behavior between start and end calibration allows one to back-interpolate the reaction data after the end of the reaction.

However, the reactor/jacket interlayer is prone to film formation and precipitation of reaction products or byproducts, and the viscosity of the reaction mixture may change dramatically in a short time interval, all of which can alter k_Q by orders of magnitude at a certain time (not continuously). Hence, interpolation is inaccurate, and even pre- and post-calibration of k_Q by separate experiments may lead to erroneous results.

Absolute Heat Flux Calibration. Absolute heat flux calorimetry (ChemiSens AB, Lund, Sweden) relies on the fact that the thermal conductivity λ_Q within the reactor wall is constant throughout the reaction, in contrast to the overall heat transfer coefficient k_Q through the wall (wall + interlayers). Calorimeter calibration can be carried out by utilizing the temperature gradient $(T_1 - T_2)/d$ (see Fig. 3). Thus, the heat flux can be calculated according to

$$k_Q \cdot A_{\mathrm{R}}(T_{\mathrm{J}} - T_{\mathrm{R}}) = \frac{\lambda_{\mathrm{Q}}}{d} A_{\mathrm{R}}(T_1 - T_2) \tag{11}$$

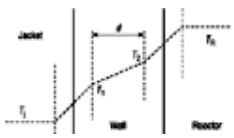

Figure 3. Absolute Heat Flux Calorimetry: Schematic of reactor wall temperature gradient (reproduced by permission of ChemiSens AB, Lund, Sweden)

Note that the immediate vicinity of the reactor wall to both sides (i.e., reaction mixture and jacket liquid) may behave like an additional thin film characterized by other distinct heat transfer resistances.

Heat Balance Calorimetry. An alternative approach for measuring the heat generated within the reactor is to determine the total heat balance of the cooling/heating liquid circulating through the jacket:

$$\dot{Q}_{\mathrm{trans}} = \dot{m}_{\mathrm{liq}} c_{p,\mathrm{liq}} (T_{\mathrm{J,in}} - T_{\mathrm{J,out}}) \tag{12}$$

where $c_{p,\mathrm{liq}}$ is the specific heat capacity of the jacket liquid. In this case, Q_{trans} is independent of both heat exchange area and heat transfer coefficient. However, it depends on the circulation rate $\dot{m}_{\mathrm{liq}}$ of the jacket liquid. Heat losses at the outer jacket wall directly influence the heat balance of the circulating liquid.

Heat balance calorimetry is relatively slow and insensitive compared to heat flux calorimetry because of the small temperature difference $T_{\mathrm{J,in}} - T_{\mathrm{J,out}}$. This also leads to stricter requirements with respect to the accuracy of temperature measurements. To achieve 0.5 W resolution, it is necessary to measure the temperature with an accuracy of about 0.001 K (compared to 0.05 K in heat flux calorimetry for the same resolution). Under typical conditions, energy resolution in heat balance calorimetry is on the order of 2 W, compared with 0.5 W in heat flux calorimetry.

Heat Compensation Calorimetry. In heat compensation calorimetry, both the vessel temperature T_R and the jacket-fluid temperature T_J are fixed with $T_J < T_R$. Heat from an additional electrical heater Q_{el} is supplied to the mixture to maintain the temperature T_R of the reactants. After the reaction has been started and a heat Q_r released, the electrical power Q_{el} is reduced to hold T_R constant. This reduction in Q_{el} is equal to the reaction heat Q_r under the condition that $k_Q A_R$ and the heat losses Q_{loss} are constant throughout the reaction.

As in heat flux calorimetry, the unknown and changing quantity $k_Q A_R$ requires calibration of the calorimeter. The hot

surface of the electrical heater is prone to the formation of films, coatings, precipitates, and hot spots.

2.2.2. *Basic Modes of Operation*

From a thermodynamic point of view, three basic modes of operation in heat flux calorimetry can be distinguished: adiabatic, isoperibolic, and isothermal calorimetry.

Adiabatic Reaction Calorimetry. An adiabatic reaction calorimeter is characterized by thermal insulation of the reaction mixture from the surroundings. As a result, the heat released during the reaction is stored within the reaction mixture. Thus, the temperature gradient of the reaction mixture directly reflects the ongoing reaction, and the total heat balance is simply given by (cf. Eq. 13):

$$\dot{Q}_r = C_{R,tot}\dot{T}_R \tag{13}$$

In practice, thermal insulation of the reaction mixture can be achieved either by means of an infinitely large thermal resistance between mixture and surroundings (e.g., Dewar calorimeter with evacuated reactor jacket) or by continuously matching the jacket temperature to that of the mixture such that there is no heat transfer through the reactor/jacket wall ($T_J(t) = T_R(t)$).

A major disadvantage of adiabatic reaction calorimetry is the interdependence of mass balance and heat balance via the reaction rate, which requires an adequate kinetic model for data analysis.

Isothermal Reaction Calorimetry. In isothermal reaction calorimetry, the reaction mixture temperature T_R is held constant, for example, by adjusting the temperature of the jacket fluid T_J or by controlling the power of a compensation heater. In practice, it is very difficult to keep the reactor strictly isothermal because of the nonzero heat transfer resistance of the reactor wall and deviations and delays in power control. Thus, heat accumulation must be taken into account in the isothermal heat balance

$$\dot{Q}_r = k_Q A_R(T_R - \bar{T}_J) + C_{p,total}\dot{T}_R \tag{14}$$

Isoperibolic reaction calorimetry is opposite to isothermal calorimetry in the sense that the jacket temperature T_J is held constant instead of the reactor temperature T_R. As a consequence, T_R changes during the reaction. Heat is both accumulated in the reaction mixture and transferred to the jacket liquid. Hence, the heat balance can be described as in isothermal calorimetry (see Eq. 14).

2.3. Applications and Instrumentation

The heat of reaction $Q_r(t)$ can be calculated according to Equation 1 if all contributions Q_i to the overall heat balance of the reactor are measured or known. If only one exothermic chemical reaction proceeds in the vessel, the reaction rate is given by Equation 2, and the actual relative calorimetric conversion $X_r(t)$ is given by Equation 15

$$X_r(t) = \frac{\int_0^t Q_r(t)\mathrm{d}t}{\int_0^{t,\mathrm{final}} Q_r(t)\mathrm{d}t} \tag{15}$$

On-line calorimetry has been used by chemical companies to monitor the extent of reactions for several decades [52-57]. Industrial-scale reactors have a relatively low ratio of heat-exchange surface area to volume compared to bench-scale reactors and calorimeters; hence, their heat-removal capacity is limited. To achieve a maximum production rate in industrial-scale reactors, the reaction rate should be matched to the heat-removal capacity throughout the reaction. On-line reaction calorimetry provides some of the data necessary for a desired reaction run and safe reactor operation (e.g., heat release rates, heat losses, reaction rate, conversion).

Reaction calorimetry has been used to monitor polymerization reactions and crystallization processes [58-60]. Strategies for on-line control and optimization of reactions on the basis of calorimetric measurements during the course of the reaction have been described [61-67]. Reaction calorimetry is the method of choice for studying thermal runaway reactions [68] (see ⟶ Plant and Process Safety). The application of calorimetry to investigating the kinetics of reactive blending and reactive extrusion is discussed in [69], and the application of adiabatic temperature rise for monitoring reaction injection molding (RIM) in [70]. On-line calorimetric data can be used to predict final product properties by means of multivariate statistics [71] or neural networks [72].

Most instrument suppliers (e.g., Mettler-Toledo, ChemiSens, HEL) offer computer-controlled calorimeters that can be run in different modes (Figure 4). The nonselectivity of reaction calorimetry in monitoring complex reactions (e.g., reactions in which parallel processes occur, copolymerizations, etc.) can be overcome by combining calorimetry with other on-line methods to obtain additional information. These methods include density, pressure, gas chromatography, dielectric spectroscopy, conductometry [73], and infrared [74], [75] and Raman spectroscopy [76] (Fig. 5; ⟶ Infrared and Raman Spectroscopy).

Figure 4. Reaction calorimeter RM-2S for heat flow and heat balance measurements (reproduced by permission of

ChemiSens AB, Lund, Sweden)

Figure 5. Reaction calorimeter RC1 equipped with React IR Infrared Spectrometer (ASI Applied Systems) for simultaneous on-line calorimetry and infrared analysis (courtesy of Mettler-Toledo GmbH, Analytical)

Calorimetric data of chemical reactions have also been determined by differential scanning calorimetry [77-80] (see ⟶ Plastics, Analysis; ⟶ Plastics, Properties and Testing).

[Top of Page]

3. Ultrasonic Methods

3.1. Introduction

The propagation of ultrasound in matter is a physical effect that is receiving increasing attention for on-line monitoring of chemical processes. Ultrasonic methods are easy to use, safe, nondestructive, and noninvasive. Signal response times on the order of milliseconds allow real-time monitoring of fast reactions.

In most applications, ultrasound is generated by piezoelectric transducers. Because of the reversibility of the piezoelectric effect, the transducers can act both as emitters and receivers of ultrasonic waves. Depending on the requirements of the application, a wide variety of readily available and cost effective piezoelectric materials can be utilized as transducers (e.g., quartz, lithium niobate, barium titanate, poly(vinylidene fluoride), zinc oxide).

Ultrasonic sensors and ultrasonic systems for materials characterization, process monitoring, and applications in chemistry are reviewed in [81-88].

3.2. Theory

A planar elastic ultrasonic wave can be described by its complex alternating pressure $p(x,t)$ as a function of time t and distance x (Eq. 16)

$$p(x,t) = p_0 \exp\left[i\omega\left(t - \frac{x}{u}\right)\right] \cdot \exp\left[-\alpha x\right] \tag{16}$$

where $\omega = 2\pi f$ (f = frequency), p_0 is the pressure amplitude, and u and α are the velocity and attenuation of sound in matter. The frequency of ultrasound waves is in the range of 20 kHz to about 1 GHz with corresponding wavelengths on the order of 1.6 cm to 0.3 µm in air ($u \approx 330$ m/s), 6 cm to 1.2 µm in liquids ($u \approx 1200$ m/s), and 20 cm to 4 µm in solids ($u \approx 4000$ m/s).

Both u and α strongly depend on the material's state and properties. In solids, the longitudinal and transverse sound velocities $u_{l,sol}$ and $u_{tr,sol}$ are related to the bulk modulus K, the shear modulus G, and the density ϱ according to

$$u_{l,sol}^2 = \frac{K + \frac{4}{3}G}{\varrho} \text{ and } u_{tr,sol}^2 = \frac{G}{\varrho} \tag{17}$$

The longitudinal ultrasonic velocity in liquids $u_{l,liq}$ depends on both the adiabatic compressibility β of the liquid and its density (Eq. 18).

$$u_{l,liq}^2 = \frac{1}{\beta\varrho} \tag{18}$$

Ultrasonic shear waves in liquids ($u_{tr,liq}$) do not propagate significantly. They are strongly dampened, and therefore not of technical importance.

The sound attenuation α of solids and liquids depends on several loss mechanisms. Intrinsic, viscous, and thermal losses may contribute to α, as well as scattering and reflection effects at interfaces in heterogeneous samples, relaxation processes in polymers, electrokinetic losses in disperse systems, and structural losses in concentrated systems (e.g.,

pseudoplasticity). Thus, α is influenced by material properties such as density, viscosity, thermal conductivity, expansion, and capacity; by the morphology of the sample (grain size and shape); and by temperature and pressure. It is difficult to derive analytical expressions for α because of its complex dependence on these parameters. However, in simple homogeneous liquid systems, α can be described as

$$\frac{\alpha_{\mathrm{liq}}}{f^2} = \frac{2 \cdot \pi^2}{\varrho \cdot u^3}\left(\frac{4}{3}\eta_s + \eta_v\right) \tag{19}$$

where η_s and η_v are the shear and the volume viscosities of the liquid. Sound is strongly absorbed at high frequencies (f^2 dependence) and cannot penetrate deeply into the material.

A third acoustic parameter of importance with respect to reaction monitoring is the acoustic impedance Z of the material (Eq. 20).

$$Z = \varrho \cdot u \tag{20}$$

It is typically measured in reflection experiments at interfaces by means of the reflection coefficient R (Eq. 21).

$$R = \frac{Z_1 - Z_2}{Z_1 + Z_2} = 1 - T \tag{21}$$

where Z_1 and Z_2 are the acoustic impedances of the materials on both sides of an interface, and T is the transmission coefficient.

A common method for determining acoustic properties is to monitor (compressional) sound pulses travelling over a well-defined distance (Fig. 6). The emitting transducer E generates a sound burst of initial amplitude p_0 (proportional to the applied voltage U_0) which travels a distance x in a time Δt. In multiple-echo mode, Δt is the time delay between successive echoes, and x is twice the distance between transducers. The dampened (and broadened) pulse reaching the receiving transducer R generates an electrical voltage U_x proportional to the pressure amplitude p_x at the receiver. By using this method, sound velocity and attenuation can simply be calculated according to Equations 22 and 23.

$$u = \frac{x}{\Delta t} \tag{22}$$

$$\alpha = \frac{1}{x} \ln \frac{p_0}{p_x} = \frac{1}{x} \ln \frac{U_0}{U_x} \tag{23}$$

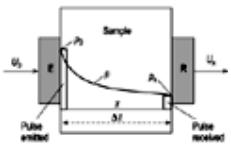

Figure 6. Schematic of ultrasonic pulse travelling measurements. (The transducer arrangement can also be used in multiple reflection echo mode.)

Measurements of pulse travelling time are highly accurate and relatively easy to carry out compared to sound attenuation measurements.

The frequency of ultrasonic sensors for reaction monitoring is generally in the 1 – 10 MHz range. In principle, the accuracy of sound velocity measurements is poor for very low frequencies (< 100 kHz) and increases with increasing frequency. However, attenuation does also increase for $f \gg 10$ MHz, and the distance between transducers must be shortened for the signal to reach the receiver. As a result, the accuracy of pulse travelling time measurements also deteriorates at very high frequencies. Typical demands on the accuracy of ultrasonic sensors for reaction monitoring are

$$\frac{\Delta u}{u} \leq 10^{-3} \quad \frac{\Delta \alpha}{\alpha} \approx 0.5\,\% \quad \frac{\Delta Z}{Z} \leq 10^{-3}$$

These conditions imply time measurements in the nanosecond and picosecond range, and of amplitude resolutions of 12 bit and higher [81]. Additionally, temperature control has to be better than ± 0.1 K.

In addition to u and α, other sound wave characteristics (e.g., wave phase φ, resonance frequency f_{res} and shift Δf_{res}) can be measured in certain ultrasonic sensor applications for materials characterization and chemical processes monitoring [81].

3.3. Applications

The basic condition for monitoring chemical processes by means of ultrasound is that physical properties that determine u, α, or Z change during the course of the reaction (see Eqs. 17 – 20). In polymer production and modification, curing and cross-linking reactions, mechanical moduli as well as the mixture's density, compressibility, viscosity, thermal conductivity, etc. alter dramatically, thus influencing sound attenuation and velocity (see Table 3). Hence, those processes are ideal for application of ultrasound to detect the extent of reaction. Even electrochemical processes can be monitored because of the sensitivity of ultrasound to the type of ion and concentration.

Table 3. Ultrasound velocities of monomers and polymers at 20 °C

Substance	Monomer	Polymer
Butyl acrylate	1233 m/s	1375 m/s
Styrene	1354 m/s	2120 m/s
Vinyl acetate	1150 m/s	1853 m/s
Vinyl chloride	897 m/s	2260 m/s

A major advantage of ultrasound (velocity) measurements in industrial applications is the capability to plug the sensor directly into the reacting mixture. They do not necessarily require a sampling loop. Care should be taken because plug-in transducers may cause fouling at their edges or lead to formation of coagula and films. The disturbing influence of gas bubbles in the reaction mixture between the transducers is demonstrated in Figure 7. Monitoring of acoustic impedance *Z* via reflection coefficient measurements may reduce the bubble disturbances. However, so far, no application of *Z* determination in reaction monitoring has been described in the literature.

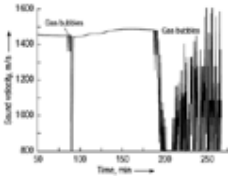

Figure 7. Influence of gas bubbles on ultrasound velocity measurements during emulsion polymerization

A simple empirical but useful approach for describing the sound velocity of multicomponent reaction mixtures is based on the assumption that each component *i* of the mixture contributes independently to the sound velocity of the mixture u_{mix} according to its concentration c_i

$$u_{mix} = u_0 + \sum_i \left(\frac{\Delta u}{\Delta c}\right)_i c_i \tag{24}$$

where u_0 is the sound velocity of a reference component (e.g., solvent, dispersion medium) whose ultrasonic velocity is either constant throughout the process or changes in a known manner, and $\Delta u/\Delta c$ is the (temperature-dependent) sound velocity – concentration coefficient, which itself may depend on concentration. However, higher order terms are neglected in Equation 24. The influence of component *i* on the sound velocity of component *j* is also neglected. Equation 24 is the basis for on-line concentration measurements in numerous applications in chemistry, biochemistry, and the foods industry.

Ultrasound has been used to monitor chemical reactions since the 1970s [89-92]. Ultrasonic on-line monitoring of polymerization reactions was reported in [93-99]. Crosslinking reactions (curing, gelation) were investigated with ultrasound in [92], [100-106]. As an example, the influence of the extent of reaction of the epoxide ring in a system consisting of the diglycidyl ether of bisphenol A (DGEBA) and 4,4′-diaminodiphenylmethane (DDM) on the ultrasonic velocity and relative ultrasonic absorption is shown in Figure 8. Viscoelastic properties of polymer melts in extruders can be monitored by in-line determination of ultrasonic parameters, along with temperature and pressure measurements [69], [91], [107-111].

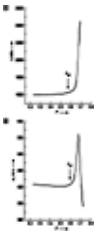

Figure 8. Ultrasonic velocity (A) and relative ultrasonic absorption (B), both measured at 1 MHz, as a function of the extent of epoxide ring reaction *P* for a DGEBA/DDM system at 70 °C; gel point P_{gel} = 0.58

(reprint from [104], with permission from Elsevier Science)

Ultrasound is sensitive to the local concentration of oil in emulsions and thus enables monitoring of creaming and flocculation processes [112], [113]. It is applicable to both dilute and concentrated emulsions.

Ultrasound is not only a tool for monitoring chemical reactions but also one for initiating them. The origin of the enhancement of chemical reactions by ultrasonication is the intensity of the cavitation bubbles produced in the sonicated medium at higher sound intensities [87], [88], [114-116] (see ⟶ Sonochemistry).

The dependence of ultrasonic parameters on grain sizes in heterogeneous media enables one to determine particle sizes and particle size distributions in latices, emulsions, suspensions, etc. even at very high concentrations of the disperse phase [117-121]. Thus, acoustic spectroscopy (as well as electroacoustic spectroscopy [119], [122]) is developing rapidly as an alternative to light-scattering methods for particle size analysis in dense media [123-127].

Ultrasonic microsensors are used for materials characterization and reaction monitoring [81], [83]. Ultrasonic microsensors based on the quartz crystal microbalance (QCM) principle are very sensitive for a wide variety of chemical applications.

The high precision and ease of measurement of the QCM frequency has made it a useful tool for measuring slight changes in mass on the QCM electrode surface. By coating the surface with a sensitive layer, the QCM can be used as selective gas or even under-liquid sensor. The determination of contact angles and surface tensions with a QCM is described in [128]. Detergency processes have been monitored in real-time by using a QCM [129]. The use of shear-horizontal (SH) acoustic plate mode (APM) devices to monitor cross-linking reactions in polymer films in real time is described in [130]. Photo-cross-linking of diacetylene thiol-based monolayers on surface acoustic wave sensors (SAW) has been monitored in situ [131]. Recent developments include the application of fiber-optic interferometric ultrasound sensors for monitoring the cure of epoxy resins [132], [133].

In addition to the active ultrasound measurements discussed above passive ultrasonic measurements can also be utilized to monitor chemical reactions. In active ultrasound, an acoustic wave is generated and changes to the wave caused by the ongoing process are being measured, whereas the acoustic emissions created by the process itself are the signals of interest in a passive ultrasonic measurement [134]. Acoustic emissions have been measured for equipment fault detection, gas – liquid mixing and flow, bubble formation and monitoring, foam break-up, powder flow and mixing, particle collision monitoring in solids and solid – fluid suspensions, compaction, and chemical reaction monitoring. In particular, chemical processes with phase changes during the course of the reaction and processes including granular material are acoustically active.

[Top of Page]

4. Dielectric Spectroscopy

4.1. Theory and Mechanisms

Dielectric spectroscopy deals with the electrical response of the polarization ***P*** (ω) of matter to the application of an oscillating electric field ***E*** (ω)

$$\boldsymbol{D}(\omega) = \varepsilon_o \boldsymbol{E}(\omega) + \boldsymbol{P}(\omega) = \varepsilon^*(\omega)\varepsilon_0 \boldsymbol{E}(\omega) \tag{25}$$

where ***D*** is the dielectric displacement vector, ω is the angular frequency, and ε* and ε_0 are the dielectric permittivity and the permittivity of free space, respectively. Several atomic and molecular mechanisms contribute to the dielectric behavior of matter [135-138]:

1. Electronic polarization (i.e., induced dipole moment due to the displacement of electrons from their equilibrium position with respect to the nucleus)
2. Displacement polarization (atomic, ionic, lattice polarization; e.g., induced dipole moment due to the deflection of ions from their equilibrium positions in molecules)
3. Orientation polarization of permanent dipoles present in the system (i.e., alignment to the direction of the external field)
4. Motion of intrinsic and extrinsic charge carriers (ions, particles)
5. Charge transfer and polarization at electrodes
6. Interfacial polarization and distortion of the electrical double layers at interfaces in heterogeneous systems

A certain mechanism contributes to the overall polarization of the sample if its characteristic relaxation time is faster than the time scale of the external field. Hence, the various polarization mechanisms can be distinguished on the basis of their different frequency dependencies.

Time scales for resonances of electronic and displacement polarization of atoms and ions are in the femtosecond (10^{-14} to 10^{-15} s; UV region) and picosecond range (10^{-12} to 10^{-13} s; IR region), respectively. Thus, these induced dipoles contribute instantaneously to the dielectric properties of matter typically measured in the frequency range 10^{-5} to 10^{11} Hz.

Key phenomena determining the change of dielectric properties during the course of a chemical reaction are orientation polarization of permanent dipoles, migration of charges, and interfacial effects in heterogeneous systems.

In contrast to electronic and atomic polarization, the orientation of permanent dipoles is time-dependent and not always instantaneous. It is governed by the (micro) viscosity of the immediate surroundings. In polymeric systems, it requires cooperative motion of chains or chain sequences, too. Dipole relaxation is the strongly temperature dependent reorientation of dipoles due to thermal fluctuations after the removal of the external field. As the frequency is scanned, various characteristic dipole relaxations accompanied by corresponding losses are observed.

The migration of charges mainly influences the dielectric losses. The (ionic) conductivity contribution to the loss factor is inversely proportional to the frequency of the external field. Ionic conductivity can be correlated with viscosity, since fluidity is indicated by the ease with which charge carriers can migrate through the sample.

The complex dielectric permittivity ε*(ω) is composed of two parts, a real (or storage) component ε′(ω) in phase, and an imaginary (or loss) component ε″(ω) out of phase with the applied oscillating electric field (Eq. 26).

$$\varepsilon^*(\omega) = \varepsilon'(\omega) + i\varepsilon''(\omega) \tag{26}$$

where *i* is $\sqrt{(-1)}$. ε ′ (ω) is proportional to the capacitance of the sample, and ε″(ω) is influenced both by the material's conductivity and by dipolar losses. A sigmoidal change in ε′(ω) and the appearance of a peak in ε″(ω) is indicative of a relaxation process (Fig. 9)

$$\varepsilon'(\omega) = \varepsilon'_0 - \frac{(\varepsilon'_0 - \varepsilon'_\infty)\omega^2\tau^2}{1+\omega^2\tau^2} \tag{27}$$

$$\varepsilon''(\omega) = \frac{(\varepsilon'_0 - \varepsilon'_\infty)\omega\tau}{1+\omega^2\tau^2} \tag{28}$$

where τ is the relaxation time. The loss factor tan δ is defined by

$$\tan\delta(\omega) = \frac{\varepsilon''(\omega)}{\varepsilon'(\omega)} \tag{29}$$

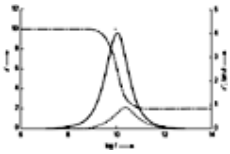

Figure 9. Dielectric relaxation: Dielectric permittivity and dielectric loss as a function of frequency

As indicated above, the dielectric permittivity is not a constant but a function of frequency. Additionally, ε^* is strongly influenced by temperature since both the transitional mobility of charge carriers as well as the reorientational motion of dipoles are temperature-dependent. For some cases in the absence of chemical reactions, the temperature dependence of dielectric properties can be described in terms of Arrhenius and Williams – Landel – Ferry (Vogel – Fulcher) type relations, depending on the nature of the dipolar motions (single activation and cooperative process, respectively).

In addition to the complex dielectric permittivity, there are several other quantities that can be alternatively derived from dielectric spectra [139], [140]. These include the impedance and admittance (or conductivity), the dielectric modulus, and the susceptibility. Admittance spectra $Y^*(\omega)$, the inverse of impedance spectra $Z^*(\omega) = Z'(\omega) - iZ''(\omega) = (Y^*(\omega))^{-1}$ for linear response, can be calculated from permittivity data according to Equation 30.

$$\varepsilon^*(\omega) = \frac{Y^*(\omega)}{i\omega C_0} \tag{30}$$

where C_0 is the equivalent capacitance of free space.

Transformations from one formalism to another can help resolve particular aspects of the relaxation processes, but no new information can be extracted in the linear response regime of the material. The impedance spectrum $Z^*(\omega)$ is often used for discussion because dielectric data can be analyzed in terms of equivalent electrical circuits consisting of resistors and capacitors connected in parallel and series. The $Z^*(\omega)$ formalism (e.g., Z'' vs. Z' plots) is particularly helpful in separating bulk and surface phenomena (e.g., polarization in the bulk and electrode polarization).

4.2. Instrumentation and Applications

Instrumentation. According to the frequency of the external electric field, several frequency regimes for dielectric experiments can be distinguished which require different measurement principles [140-142] (Fig. 10).

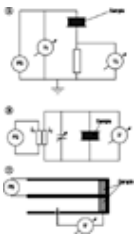

Figure 10. Principles of dielectric measurements: A) Frequency response analysis technique; B) Resonance circuit; C) standing wave in hollow conductor

PS: Power source

The common instrumentation in the *impedance regime* from about 10^{-5} to 10^{8} Hz includes current – voltage measurements in electrode/sample/electrode sandwich arrangements of electrical (ratio-arm) bridges, and resonance (oscillating) circuits. The main techniques for dielectric reaction monitoring are (1) the direct a.c. measurement of amplitude and phase between voltage and current over a certain frequency range by using a frequency response analyzer (FRA), and (2) step voltage application (or removal) with rapid acquisition of the current response over time and transformation of the signal into the frequency domain.

Minimum time intervals for dielectric measurements are on the order of tenths of seconds, so dielectric spectroscopy in the impedance frequency regime is suitable for on-line monitoring of chemical reactions. Typical accuracies of measurements achievable in the impedance frequency regime are about ± 0.05 % for ε', ± 0.2 % for ε'', and better than 10^{-4} for tan δ.

Dielectric cells for liquids are generally simple and consist of two concentric cylinders with the fluid in the space between them. These measurements are of very high precision because of the good dimensional stability of the cells. Solid samples are typically sandwiched between metal plate electrodes as in a capacitor. Compared to liquid cells, solid sample set-ups are quite complicated because they must take into account for the sample's thermal coefficient of expansion.

The *radio regime* from about 10^{6} to 10^{10} Hz requires resonance circuits (low ω) or guided-wave techniques (high ω; e.g.,

standing waves in hollow conductors).

The typical equipment for the *microwave region* from 10^9 to several 10^{11} Hz is coaxial or rectangular waveguides.

Propagating wave techniques are utilized in the *IR and UV – VIS regimes* above a few 10^{11} Hz with Fourier transform technique as the method of choice.

Applications. Chemical reactions are accessible to reaction monitoring by dielectric spectroscopy if they lead to changes in the rate of motion of molecules contributing to the dielectric response of the mixture (e.g., rotational or vibrational motion of dipoles and translational motion of charge carriers).

Many applications of dielectric spectroscopy published so far are for monitoring the cure of resins. Dielectric spectroscopy can detect changes in the polymer cross-link density during the course of polymerization (Fig. 11). When van der Waals or other weak interactions between molecules are replaced by covalent bonds during the conversion of low-viscosity monomers into polymers, the dipole moments of molecules and the mechanisms of their Brownian diffusion change, as does the size of the diffusing moieties. The restrictions on the diffusion of dipoles and charge carriers increase [143-148].

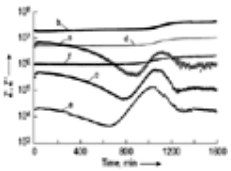

Figure 11. Real and imaginary impedance as a function of time with frequency as parameter during the reaction of the diglycidyl ether of bisphenol F (DGEBF) with 4,4′ methylenedianiline (MDA) at 36 °C [146]

a) Z' = 500 Hz; b) Z'' = 500 Hz; c) Z' = 2000 Hz; d) Z'' = 2000 Hz; e) Z' = 10 000 Hz; f) Z'' = 10 000 Hz

Both the real and the imaginary parts of the dielectric permittivity can have dipolar and ionic-charge polarization components. At lower frequencies, ε'' is typically dominated by the mobility of ions in the sample, whereas dipolar reorientation mobility contributes to ε'' at higher frequencies (Eq. 31).

$$\varepsilon^*(\omega) = \varepsilon^*_{\text{dipolar}} - i\frac{\sigma_{\text{dc}}}{\omega\varepsilon_0} \tag{31}$$

Here σ_{dc} is the d.c. conductivity due to ever-present ionic impurities [149], [150]. The ionic mobility (i.e., σ_{dc}) is a molecular probe for quantitatively monitoring viscosity changes during cure. At short curing times, ε'' is dominated by the d.c. conductivity. When gelation occurs, the d.c. conductivity decreases strongly since the mobility of the charge carriers decreases due to network formation. Peaks in ε'' indicate when dipolar reorientation processes contribute to ε''.

The ionic conductivity signal of dielectric measurement data has been used to study curing of epoxy resins. The conversion was calculated based on a Williams – Landel – Ferry model approach [151]. Dielectric spectroscopy in the 10 to 10^5 Hz frequency range with a ratio-arm bridge has been used to monitor the UV polymerization of a polymer-stabilized liquid crystal cell containing a mixture of nematic and chiral liquid crystals together with DSM monomer and photoinitiator [152]. The authors discuss their results in terms of capacitance units. They have shown that the decreasing loss serves as an excellent monitor for changing molecular processes and the state of photopolymerization, whereas the capacitance itself does not change dramatically during the course of the polymerization. Low-frequency losses observed at the beginning of the reaction are mainly caused by ionic impurities. After polymerization, the liquid crystal cell has the dielectric characteristic of a relatively low loss dielectric.

An electrical impedance spectroscopy method has been developed as a real-time online control tool for thermoset composite manufacturing by exploiting the change of the imaginary impedance maximum as a reaction progress indicator [153].

In particular, dielectric spectroscopy is the method of choice for following the curing of polar coatings. In these thin-layer systems, very high signal-to-noise ratios can be achieved because their capacitance is inversely proportional to their thickness, like the capacitance in a plate capacitor as a function of interplate distance. A special thin-layer dielectric method has been devised which measures the dissipation factor of an adhesive during the cure of adhesively bonded joints and converts it into the degree of cure of the adhesive [154].

Dielectric spectroscopy has been also applied to disperse systems [155-158]. The relaxation pattern is the dielectric fingerprint of a colloidal system [159]. Correlations between relaxation frequencies and suspension properties such as particle volume fraction, electrolyte concentration, Zeta potential, pH, surface conductivity, particle radius, Debye screening length, degree of ionization, and diffusion coefficient have been developed [160], [161]. These characteristics change during reactions in colloidal systems, and on-line monitoring is possible. Moreover, dielectric spectroscopy is especially valuable for studying concentrated and opaque dispersions. For a latex exhibiting a low-frequency relaxation process, the relaxation time at the loss maximum $\tau = 1/\omega_{max}$ is proportional to the square of the particle radius, and inversely proportional to the diffusion coefficient of the electrical double layer of counterions up to 30 % total solids [162].

[Top of Page]

5. Optical Spectroscopy

5.1. Introduction

Optical spectroscopy covers the range of the electromagnetic spectrum from the ultraviolet (UV, $\lambda \gg$ 10 nm) to the

infrared (IR, λ < 1 mm). Experimental equipment for optical spectroscopy in this wavelength range is constructed with typical optical elements such as mirrors, lenses, gratings, and optical filters. The incident energy is delivered by light sources. The long-wavelength end of the electromagnetic spectrum (i.e., millimeter range) requires completely different experimental techniques (e.g., hollow conductors for microwave spectrometry) as does the short-wavelength end ($\lambda \ll$ 100 nm) with special X-ray and γ-ray equipment.

Electronic absorptions determine the spectral features observable in the UV – VIS region. Molecular vibrations, rotations, and combined rotational-vibrational transitions of molecules can be detected in the near (NIR) to the far (FIR) infrared wavelength range. Spectra can either be recorded in absorption (UV, VIS, NIR, IR, FIR) or emission (IR, Raman, fluorescence).

For general reading, fundamentals, instrumentation and applications, see [12] and ⟶ Ultraviolet and Visible Spectroscopy and ⟶ Infrared and Raman Spectroscopy.

5.2. Instrumentation for Reaction Monitoring

Spectroscopy over the entire spectral range of optical wavelengths has been applied to on-line process monitoring for several decades. The vast majority of applications is based on nondispersive instruments, and light absorption or emission is typically measured either at a single wavelength or by integration over a certain (rather small) wavelength range.

In absorption spectroscopy, the light is generated by monochromatic light sources (including lasers) or polychromatic light sources combined with optical filters. Tuning of the wavelength to characteristic absorptions of certain molecules enables one to detect substances and to observe the variation of their concentration as a function of time with high selectivity and sensitivity.

In transmission measurements, the fundamental relation between the incident light intensity I_0, the transmitted light intensity I_l, and the concentration of the analyte c in a cell of optical length l is given by the Lambert – Beer law:

$$\log \frac{I_0(\nu)}{I_l(\nu)} = \log \frac{1}{\tau(\nu)} = A(\nu) = \varepsilon(\nu) cl \qquad (32)$$

where $\varepsilon(\nu)$ is the (decadic) extinction coefficient of the sample at frequency ν, l is the length of the measurement cell, A is the absorbance (extinction), and τ is the transmittance of the sample. Monitoring procedures based on this classical photometric method are still extensively used because of its simplicity, ruggedness, sensitivity, and low cost and maintenance requirements. It works best for well-characterized continuous material streams but is less advantageous for dynamic processes (e.g., some batch operations).

The number of substances that can be detected simultaneously by nondispersive instruments in absorption and emission modes is limited, and detailed molecular information on the species undergoing a reaction is not available. These restrictions can be overcome by parallel installation of several nondispersive instruments operating at different wavelengths, or utilization of dispersive instruments (spectrometers) to exploit the whole spectral information.

With classical dispersive wavelength-scan spectrometers, data collection is time-consuming because such an instrument records the data points of the spectrum consecutively. Therefore, dispersive scanning spectrometers are not suitable for process monitoring.

However, the substitution of point detectors by array detectors in dispersive instruments offers the opportunity to acquire large amounts of spectrally and/or spatially resolved data in extremely short periods of time. Semiconductor array detectors are characterized by high quantum efficiency, wide dynamic range, fast response, low noise, low power requirement, and continuously decreasing costs. Array detectors suffer from either limited resolution or limited bandwidth. However, the two-dimensionality of some types of array detectors in combination with cross-dispersion enables the complete spectrum to be obtained with high resolution. Since the whole spectrum is available with one shot of light on the detector, chemical reaction monitoring on the sub-second timescale is possible. Another advantage of two-dimensional array detectors is the ability to record several spectra simultaneously in different lines of the array thus enabling multiplexing. Fundamentals and applications of array detectors [charge-transfer devices (CCD, CID), photodiode arrays, and photoconductor arrays] in both absorption and emission spectroscopy are reviewed in [163] (see ⟶ Ultraviolet and Visible Spectroscopy – Detectors).

With Fourier transform (FT) spectrometers, the spectral information is first recorded in the time domain as an interferogram and then Fourier transformed into the frequency (wavenumber) domain. The optical throughput of an FT instrument is very high because light of all frequencies is recorded simultaneously. Depending on the spectral resolution and the signal-to-noise ratio required, FT spectrometers are suitable for monitoring very fast chemical processes on the timescale of seconds to minutes. However, they are less rugged than photometers or array-detector instruments because the interferometer is susceptible to vibrations, for instance, in a production-plant environment.

The introduction of fiber optics has been a major step forward in reaction monitoring by optical spectroscopy. Nowadays, most instrument suppliers offer their spectrometers with optional fiber optics. Optical fibers connect the remote instrument with the reactor or process line for on-line monitoring with the following advantages:

- Reduced risk of fire or explosion in production units
- Reduced interaction of the delicate spectrometer with the hostile process environment
- Enhanced accessibility of various measurement points in a process stream or reactor
- Simplified multiplexing

Light may be delivered and collected through separate single fibers or fiber arrays; alternatively, fibers can be arranged in

integrated fiber bundles in which a light source fiber is surrounded by detector fibers, or in which source and detection fibers are randomly distributed within the bundle. A review of the developments in fiber optic applications in molecular spectroscopy is given in [164] (see ⟶ Chemical and Biochemical Sensors). Commonly used fiber types are listed in Table 4.

Table 4. Fiber types and typical optical wavelength range [164]

Fiber type	Approximate optical range
High-purity fused silica	200 – 2000 nm
Plastics	350 – 1100 nm
Glass	350 – 1800 nm
Low-OH fused silica	250 – 2700 nm
Zirconium fluoride	2 – 5.5 µm
Chalcogenide	3 – 11 µm
Silver halides	5 – 20 µm

Recently, the development of fiber-optic chemical sensors (FOCS) based on evanescent wave absorption (EWA) spectroscopy has experienced much progress and improvement. FOCSs can be employed for real-time in situ monitoring by simply immersing them in the mixture and establishing a connection between the fiber, the light source, and the detection system. They offer fast response times on the order of seconds with all the advantages of attenuated total reflection (ATR) spectroscopy, and considerably improved sensitivity.

Figure 12 shows various instrument/probe/process vessel configurations for the application of optical spectroscopy to the monitoring of chemical reactions. Spectra can either be recorded in transmission, reflection, transflection, or attenuated total reflection. Probes can be directly immersed into the reaction mixture, or the spectra can be recorded through windows in the reactor or pipe wall.

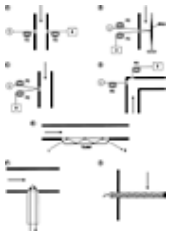

Figure 12. Instrument probe process vessel configurations for the application of optical spectroscopy to monitoring chemical reactions

A) Transmission; B) Transflection; C) Reflection; D) Scattering (90 °); E) Attenuated total reflection (crystal); F) Attenuated total reflection (single-bounce tip); G) Evanescent-wave fiber optic

FO = fiber optic; L = light source; D = detector

The polarization state of reflected light can be determined in addition to its intensity with spectroscopic ellipsometry methods ranging from UV to IR wavelengths [165]. Spectroscopic ellipsometry is a powerful tool to characterize optical properties of thin films and near-surface regions of bulk materials. It has been applied to study the deposition of material in CVD processes, thermal evaporation or magnetron sputtering [166], [167].

5.3. Applications of Optical Spectroscopy

5.3.1. *UV – VIS Spectroscopy*

UV – VIS spectroscopy is highly sensitive, but the spectra are less informative because they typically consist of a few broad absorption peaks. Therefore, UV – VIS spectroscopic reaction monitoring is less commonly used than other spectroscopic techniques. Typical applications of UV – VIS spectroscopy include photometric gas analysis in transmission gas cells for the detection of sulfur dioxide, nitrogen oxides, chlorine, mercury vapor, toluene, benzene, methane, ethanol, etc. Detection limits are in the order of tens of ppm in air [8].

A major advantage of the UV – VIS spectral region is that water does not absorb considerably in the entire VIS region up to the mid-UV, thus making it attractive for water and wastewater analysis. UV – VIS absorption measurements in aqueous or other fluid process streams are typically performed in transmission or ATR modes. With ATR probes, even highly opaque mixtures and liquid systems containing strongly scattering particulate solids are accessible to process analysis.

In disperse systems, light scattering by particles accounts for much of the attenuation of the incident light. Thus particle size measurement by UV – VIS scattering methods is common practice in colloid laboratories and at production sites.

The general applicability of UV – VIS spectroscopic techniques as well as their advantages compared to engineering process data in order to obtain characteristics of ongoing chemical reactions, to control reactions, to predict product properties, and to determine end points of batch and blending processes has been discussed in conjunction with data handling methods in detail in [168].

The versatility of UV – VIS spectroscopy in process monitoring has been extended by the development of UV array spectrometers [169].

Optical fibers for use in the VIS region are readily available. Transmission of visible light through optical cables is possible

over long distances for remote process monitoring. It has been shown that fiber-optic UV – VIS spectroscopy and chemometric data treatment (window factor analysis) enable one to detect reaction intermediates of an industrially significant process that could not be detected otherwise (e.g., by HPLC) [170]. UV – VIS fiber optics have also been applied to monitor chemical processes in supercritical fluids [171]. In the short-wavelength UV, however, the optical absorption of the fiber core and/or cladding material restricts light throughput. Recently, evanescent-wave fiber optic absorption sensors for use in the UV and VIS regions have been described [172], [173]. UV reflection spectroscopy with a fiber optic accessory has been used to characterize the cure of bis(maleimide) resins [174], [175].

5.3.2. *NIR Spectroscopy*

Absorption bands in the near infrared originate from electronic transitions of the heaviest atoms and from overtones and combinations of fundamental vibrations in the IR. In organic substances, C – H, N – H, and O – H vibrations give rise to dominant absorptions in the near-IR. Therefore, the detection of water was one of the first on-line applications of NIR spectroscopy. However, water interference may also complicate NIR spectroscopy of aqueous samples [176]. NIR-active vibration overtones and combinations are characterized by low molar absorptivities and high scattering efficiencies, and the NIR bands are broad and overlapped.

The peculiarities of NIR spectrometry lead to important consequences in practice:

1. Near IR wavelengths are able to penetrate into and sometimes even through a sample, thus enabling measurements on real world samples in the millimeter or centimeter thickness range. Spectra can be recorded on most types of liquid, solid, powder, or gaseous samples without sample pretreatment. Sample size may vary from tiny bits to large chunks.
2. NIR spectra contain all the spectral information necessary for fingerprinting, albeit buried under broad and strongly overlapped absorption features. In most cases, the spectra cannot be interpreted directly, unlike the individual peaks of IR spectra, and calibrations and chemometric methods have to be exploited for quantitative analysis. This may cause problems in NIR reaction monitoring of batch processes with frequently changing recipes and reaction regimes because of the necessity to provide an extensive training data set for each process [177].

Typical analysis time is on the order of seconds or even shorter, with sensitivities down to tenths of a percent; hence, NIR is an ideal tool for observing fast reactions. Spectra of chemically changing mixtures can be recorded through quartz glass windows mounted in reactors and pipes or by fiber optic immersion probes for diffuse reflectance, ATR, transmission, and transflection measurements. The application of cheap silica fibers allows truly remote sensing with distances up to kilometers between process and spectrometer because of the extremely low attenuation of quartz in the VIS to NIR range.

NIR equipment suitable for reaction monitoring includes both holographic grating/semiconductor array detector instruments and Fourier transform spectrometers. Light-emitting diode (LED) based instruments have no moving parts because each wavelength band is produced by a different diode with precise wavelength tunability and monochromaticity, thus making the instrument rugged to harsh process environments. LED spectrometers are highly sensitive and selective, with fast response and exceptional dynamic range. So far, they have been applied to on-line monitoring of HF and H_2S traces in refineries. However, their high costs are prohibitive for most on-line applications. Acousto-optic tunable filter (AOTF) based instruments have been occasionally used for on-line analysis. Hadamard transform (HT) spectrometers have not been described for NIR on-line applications so far.

Typical problems with NIR measurements are the need to check the calibration of the instrument, since the instrumental response may change after some time [178], and the sensitivity of NIR spectra to the sample temperature, which complicates training procedures. NIR equipment and the application of NIR to reaction monitoring are reviewed in [179-183] (see ⟶ Infrared and Raman Spectroscopy).

NIR spectroscopy has been widely used for the characterization of epoxy curing [184, 185] with conventional fiber optics [186], [187], evanescent wave high-index fiber optic sensors [188], and the NIR ATR technique [189]. Other fiber optic applications include reaction monitoring in explorative organic synthesis [190], real-time analysis of light alkenes in high-pressure reactors at elevated temperatures [191], remote monitoring of styrene emulsion polymerization in the short-wavelength NIR region [192], and determination of octane in refinery processes [193]. A fiber optic coupled AOTF NIR spectrometer has been described for dryer effluent monitoring in a chemical plant [194]. The number of papers describing the utilization of NIR spectroscopy to monitor the progress of polymerization reactions of silanes [195] and vinyl monomers [196-202] is continuously increasing. One of the major advantages of NIR spectroscopy in polymerization monitoring is the capability to extract molecular mass information for reaction control instantaneously that cannot be measured otherwise in a timely manner [203-205]. NIR spectroscopy has been shown to be suitable for on-line detection of reaction completion in a closed-loop hydrogenator in pharmaceutical production [206], for end-point detection in pharmaceutical high-shear granulation processes [207], and in-line monitoring of the moisture content in fluid-bed dryers [208]. Diffuse reflectance NIR spectroscopy has been evaluated as an on-line technique for the monitoring of powder blending [209-211]. Gas-phase real-time monitoring by transmission NIR spectroscopy is described in [212], [213]. NIR spectroscopy has been utilized to measure major components in the production of acetic acid [214]. On-line particle size monitoring based on NIR spectroscopy has been described for pharmaceutical nanoparticles [215], [216] as well as for emulsion polymers [217]. A promising application of remote-sensing NIR spectroscopy is the on-line identification and classification of plastics material [218-220].

Near-infrared spectroscopy is one of the few analytical methods applicable to real-time monitoring of chemical processes in polymer melts during extrusion [221]. Figure 13 shows a ceramic transmission NIR probe with variable optical pathlength that is adapted to a corotating twin-screw extruder and interfaced to an NIR spectrometer via fiber optic cables. On-line NIR spectra, as shown in Figure 14, provide the quantitative composition of a polypropylene/ethylene – vinyl acetate copolymer mixture during blend formation with an accuracy of about 1 %. With the help of ATR (Fig. 15) and diffuse reflectance probes the formation of styrene – maleimide copolymer (Fig. 16), and the filler concentration in molten, opaque polypropylene/chalk powder composites have been monitored. In the latter application, the inhomogeneous

distribution of filler particles may lead to substantially different spectra during the course of extrusion. The scattering contribution of the particles to the NIR spectra has to be corrected before chemometric analysis. Appropriate chemometric calibrations models are required in both cases [222-224]. HANSEN et al. [225-228] were even able to derive rheological properties of ethylene – vinyl acetate copolymers from NIR spectra recorded in-line in an extruder via fiber optics.

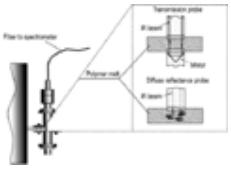

Figure 13. Ceramic transmission NIR probe with variable optical pathlength and diffuse reflectance dipper probe for insertion into extruder (courtesy of IPF Dresden)

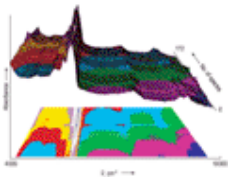

Figure 14. On-line NIR spectra of a polypropylene/ethylene – vinyl acetate copolymer during blend formation (courtesy of IPF Dresden)

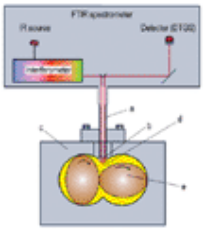

Figure 15. ATR immersion probe adapted to a corotation twin-screw extruder (courtesy of Carl Hanser Verlag)

a) ATR immersion probe; b) ZnSe crystal; c) Intermediate plate; d) Melt in the intermeshing zone; e) screw

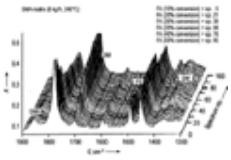

Figure 16. Formation of styrene – maleimide copolymer monitored with a diffuse reflectance dipper probe (courtesy of Carl Hanser Verlag)

FA = C-18 fatty amine; PS = polystyrene; SMA = styrene – maleic anhydride copolymer; SMI = styrene – maleimide copolymer

5.3.3. *IR Spectroscopy*

Infrared spectroscopy has a long tradition of use in chemistry, materials science, physics, and even reaction monitoring [229]. It is one of the standard techniques in most analytical laboratories. The IR region of the spectrum is highly specific and very sensitive to molecular structures. Individual peaks can be assigned unequivocally to chemical species and calibrated for analyte concentration. The timescale for FTIR spectroscopic analysis is on the order of seconds. All of this makes IR spectroscopy attractive for on-line process monitoring.

High molar absorptivities in the IR are advantageous for gas analysis but may present challenges for liquid analysis because sample layer thicknesses (optical pathlengths) would have to be in the micrometer range to transmit IR radiation. This distance is too short for circulation of the reaction mixture, so transmission and transflection measurements cannot be used for process monitoring.

The technique of choice is ATR spectroscopy. The short penetration depth of the evanescent wave is both the strength and a limitation of ATR spectroscopy. ATR crystals are prone to surface fouling and layer build-up, which leads to interference with the process signals. Additional issues to be considered with respect to on-line applications of ATR are the scratch resistance of the crystal; the relatively low light throughput, which requires sensitive detectors; and the linearity (or nonlinearity) of calibration.

A major obstacle to industrial IR on-line applications is the lack of appropriate fibers transparent to IR radiation. The chalcogenide fibers currently used are very expensive and have only limited light throughput and hence do not enable remote monitoring. Infrared spectrometers should be installed close to the process to allow the use of automated sampling devices, short fibers, or optical systems that transmit the light via mirrors and tubes to the analyzer. The latter option is shown in Figure 5, in which a “mirror-arm” is interfaced to an FTIR spectrometer. In Figure 17, a waterfall spectrum and the corresponding line intensities as a function of time are shown for the reaction of secondary amines with paraformaldehyde. In this case, a linear calibration model makes it possible to examine the reaction kinetics on-line on a quantitative basis.

Figure 17. Waterfall spectrum (a) and the corresponding line intensities (b) as a function of time for the reaction of

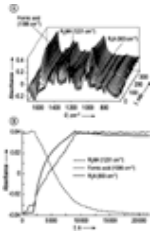

secondary amines with paraformaldehyde (courtesy of Mettler-Toledo GmbH, Analytical)

In aqueous systems, the strong water absorption in the IR region may interfere with spectral features of interest for reaction characterization. Nevertheless, on-line monitoring of 2-ethylhexyl acrylate/styrene emulsion copolymerization with an ZnSe ATR fiber optic FTIR device has been demonstrated successfully [230].

Applications of FTIR spectroscopy to polymerization reactions include kinetic studies of laser-induced photopolymerization directly in a spectrometer [231], multiacrylate polymerization studies with a quartz crystal ATR device [232], homo-, co-, and terpolymerization reaction monitoring of acrylates and vinyl acetate [233], [234], and monitoring of carbocationic polymerization with ATR probes [235], [236]. A water-cooled mid-infrared chalcogenide fiber-optic probe with an ZnSe crystal was described for in situ monitoring of an acid-catalyzed esterification reaction in toluene at 110 °C [237]. Cross-linking reactions in various resins have been studied by IR spectroscopy [238-242], and FTIR imaging with focal-plane array detection [243]. The melt crystallization process of isotactic polystyrene was studied by means of in situ FTIR spectroscopy [244].

On-line gas process stream analysis was carried out by piping the gas through a transmission gas cell [245] or by means of photoacoustic gas sensors [246]. Chemical processes in thin films and coatings have been investigated by photoacoustic FTIR spectroscopy [247] and in situ IR ellipsometry [248]. FTIR in situ spectroscopy has been applied to monitor yeast fermentation processes [249], and the homogeneously catalyzed liquid-phase hydrolysis of sucrose [250]. It has been shown that pharmaceutical batch processes can be monitored by FTIR real-time analysis [251]. On-line measured mid-IR spectra in combination with multivariate methods (orthogonal projection approach, principal component analysis, multivariate curve resolution) enable the operator to detect the end point of a chemical analysis [252]. ATR – FTIR in conjunction with multivariate quality-control methods has also been explored to control the manufacturing process of PVC extra-corporal medical disposables [253].

5.3.4. *Raman Spectroscopy*

Fundamentals. Raman spectroscopy combines some of the advantages of (FT)IR and NIR spectroscopy without the limitations associated with these techniques. Nowadays, it is one of the fastest growing areas of analytical chemistry.

Raman spectroscopy produces well-resolved spectra of fundamental vibrations comparable to IR spectroscopy. The interaction probabilities of light with the molecule for the two techniques are quite different. Infrared absorption is favored when a molecule has a permanent dipole which is modulated by the vibration. Raman scattering occurs when the molecule is polarizable, with the polarizability modulated by the vibration. Thus the intensities of IR and Raman bands of the same substance are fundamentally different, and the two methods provide complementary information about the molecule. Some of the bands can even be absent in one or the other spectrum.

The Raman line intensity is directly proportional to the number of corresponding oscillators in the scattering volume and the intensity of the illuminating radiation. Line intensity changes primarily reflect concentration changes within the sample, thus making calibration of the spectra rather simple and straightforward. Raman spectroscopy does not require extensive sample preparation, and Raman spectra can be taken from almost all samples without pretreatment.

The Raman effect is an extremely weak, inelastic scattering process. Typically, only 10^{-6} to 10^{-9} of incident photons undergo Raman scattering, and low-loss optical devices are required for proper recording of spectra. In Raman spectroscopy, there is always a competitive mechanism for light emission present, namely, fluorescence of the sample. In some cases, the quantum yield of fluorescence outweighs the Raman scattering efficiency by several orders of magnitude, and the fluorescence of the sample completely obscures the Raman spectrum. The most convenient way to avoid sample fluorescence is excitation at longer wavelengths. However, as the excitation wavelength increases, the light scattering intensity decreases according to a λ^{-4} power law.

Raman signals of some species are much more sensitive than the corresponding IR absorptions. The ability to measure Raman spectra of symmetrical diatomic molecules (for instance O_2, N_2, H_2, F_2, Cl_2, Br_2, I_2 are not active in IR absorption) opens up opportunities in gas analysis. Raman spectroscopy is sensitive to nonpolar molecular vibrations. Hence, double or triple bonds in monomeric or polymeric molecules are strong Raman scatterers. Vinyl or diene monomers can easily be identified, and their concentrations determined by means of their double bonds. Therefore, Raman spectroscopy is an ideal tool for monitoring the disappearance of monomers during the course of polymerization, and for detecting residual monomers in final products. Metal carbonyls in catalysts and cyanides in plating baths give rise to strong Raman signals due to their triple bonds. Raman spectroscopy can be used to monitor the quality of diamond or diamond-like carbon and to distinguish between isomers [254].

Water, with a strong molecular dipole, has an intense IR absorption, but a very weak Raman response. Thus water obscures IR spectra, whereas Raman scattering is more or less oblivious to the presence of water. Hence, Raman spectroscopy can easily be applied to aqueous solutions, emulsions, latices, and suspensions. Disperse systems scatter light because of refractive index differences between particles and the surrounding continuous phase (e.g., elastic Rayleigh or Mie scattering of latices, emulsions, aerosols). Multiple elastic light scattering reduces severely the scattering

volume from which the light for Raman spectroscopy is collected. Hence, confocal illumination and collection optics are essential for efficient sample excitation and signal detection. Light scattering affects spectral intensities but not band shapes. With confocal optics, interference of both the window material and films built up on window surfaces with the spectrum is much reduced, and spectra can be obtained from samples through colored or opaque sample container walls.

Instruments. Dispersive scanning Raman spectrometers operating in the UV or VIS region are not very suitable for process monitoring because of the time required to record a spectrum. Additionally, the short wavelength excitation may cause severe sample fluorescence. The main advantage of dispersive scanning instruments is the very high resolution achievable by using double or triple monochromators.

As in IR spectroscopy, FT Raman spectrometers are suitable for monitoring even very fast chemical processes on the timescale of seconds. For practical and cost reasons, most instrument manufacturers use the same type of interferometers in their IR and FT Raman spectrometers. Hence, excitation in FT Raman spectroscopy is typically at 1064 nm with Nd:YAG lasers. Therefore, sample fluorescence is very much reduced, but FT Raman spectra of aqueous samples are modified by self-absorption of water in the NIR (i.e., for wavenumbers above 2000 cm^{-1} with excitation at 1064 nm). Pros and cons of FT Raman spectroscopy are reviewed in [255-257].

The application of holographic transmission gratings and semiconductor array detectors (CCD or CID) has led to the rediscovery of dispersive Raman spectrometers. The major advantage of these instruments is they do not contain any moving parts and are thus rugged and compact for on-line applications in chemical plants. However, the insensitivity of a CCD at wavelengths above 1050 nm restricts the wavelength of the incident light. For a CCD-based instrument, excitation of the sample has to be done below 800 nm to obtain the entire spectrum up to 3000 cm^{-1}. Most instruments for reaction monitoring are equipped with recently developed inexpensive and compact NIR solid-state diode lasers operating between 780 and 850 nm, whereby 785 nm or 840 nm are commonly used. A review on diode lasers for Raman spectroscopy is given in [258].

In this wavelength range, low-cost silica optical fibers deliver the laser light to the measurement site and return the signal efficiently for remote monitoring. Holographic filters integrated into the probe head remove the Raman signal generated within the excitation optical fiber and prevent reinjection of elastically scattered laser light into the signal return fiber. Raman spectrometry with fiber-optic sampling is described in [259-262].

For trends in instrumentation and applications, see [263-267].

Applications. Prior to the 1990s, only a few publications deal with Raman reaction monitoring. The validity of the method has been demonstrated for the suspension polymerization of styrene [268] and vinyl chloride [269], the thermal polymerization of styrene [270-272] and methyl methacrylate [272], the solution polymerization of methyl methacrylate [273], the γ-initiated polymerization of diacetylene [274], and the microemulsion polymerization of styrene and methyl methacrylate [275].

The decreasing intensity of the ν(C=C) Raman lines of the monomer during the course of the reaction was monitored as a measure of the extent of conversion to polymer. The positions of the Raman double bond stretching vibration of several monomers are listed in Table 5. Raman spectra recorded during emulsion copolymerization of styrene and butyl acrylate are shown in Figure 18. The ν(C=C) double bond peaks of the two monomers cannot be resolved since the spectral resolution of the instrument was on the same order of magnitude as the difference between their peak maxima (ca. 7 cm^{-1}). The double bond peak disappears almost completely during the reaction. Based on the spectral changes as a result of polymerization, the conversion of the two monomers as a function of time can be calculated with an accuracy of about ± 1 %. The combination of Raman on-line monitoring and automated spectra evaluation enables chemical reactions to be controlled efficiently.

Table 5. Double bond Raman lines

Monomer	$\tilde{\nu}$(C=C), cm^{-1}
Vinyl chloride	1607
Acrylonitrile	1610
Vinyl sulfonate	1619
N-Methylacrylamide	1629
Styrene	1631
Methyl acrylate	1635
2-Ethylhexyl acrylate	1637
Butyl acrylate	1638
Diisopropyl fumarate	1638
Ethyl acrylate	1638
Butadiene	1639
Glycidyl methacrylate	1640
Methyl methacrylate	1641
Allyl methacrylate	1641/1648
Veo Va 9/10*	1646

Vinyl acetate	1648
Vinyl propionate	1648
Crotonic acid	1658

* Vinyl esters of versatic acids.

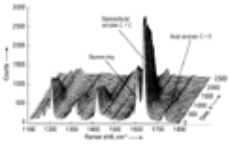

Figure 18. Raman spectra recorded during emulsion copolymerization of styrene and butyl acrylate. The spectra were recorded with a fiber-based instrument equipped with a CCD detector. Laser excitation at 785 nm with an output power of about 90 mW was performed through a glass window in the reactor wall. The accumulation time for each spectrum was about 15 s. Spectra were recorded at regular intervals of 1 min. Hence, the instrument meets the requirements for an on-line process monitoring system.

Between 1990 and 2005 a wealth of publications appeared on in situ Raman investigations, for instance, of polymerization reactions [276-285], epoxy resin cure [286-288], crystalline-state photoreactions [289], fast high-pressure decomposition reactions [290], hydrolysis of acetic anhydride to acetic acid in a hydrothermal/supercritical water reactor [291], liquid-phase chemistry of aliphatic organic peroxides [292], crystallization of polymer melts [293], a PCl_3 production process [294], petroleum distillate quality in pipelines [295], synthesis and hydrolysis of ethyl acetate [296], the manufacture of titanium dioxide [297], the biotransformation by yeast of glucose to ethanol [298], homogeneous catalysis (Heck reaction) using a solvent band as internal standard to improve quantification [299], and chlorosilane stream composition [300]. Chemical reactions have even been monitored in aerosol particles by means of Raman spectroscopy [301-304]. There is evidence that Raman spectra also contain information on the size of particles undergoing a chemical reaction [305], [306].

Raman spectrometers have been attached to microreactors for reaction monitoring [307]. Multiplexing in a chemical production environment can easily be achieved by Raman systems equipped with fiber optics [308]. The development of low-cost low-resolution spectrometers may open new on-line application opportunities in the future [309].

Micro-Raman spectroscopy has been successfully installed for process and growth monitoring and control of GaN, AlGaN, and AlN [310], and for on-line defect review in IC manufacturing processes [311]. Raman compositional mapping of paper coatings enables one to detect the spatial distribution of pigments, binders, and fillers [312]. Large-spot noncontacting Raman probes have been developed to on-line inspect pharmaceutical formulations and tablets [313].

The study of Raman spectra of adsorbed molecules on surfaces is one of the most promising areas in Raman spectroscopy. Molecules adsorbed on metal surfaces show enhancement of the scattering efficiency by up to seven orders of magnitude [314], [315]. This surface-enhanced Raman scattering (SERS) is ultrasensitive for detecting numerous adsorbed compounds by means of their vibrational spectra. Originally, SERS experiments were carried out with either bare metal surfaces or metal colloids mixed with solutions of the sample. Covering the substrates with extremely thin modifying layers enables the construction of chemical sensors for remote monitoring [316-319].

A nonlinear Raman technique, stimulated Raman scattering (SRS), has been used to monitor the bulk polymerization of styrene and methyl methacrylate [320]. In contrast to spontaneous Raman scattering, most compounds show only a few Stokes lines in their SRS spectrum. As a result, spectral interference of SRS-active compounds in a reaction mixture is minimized.

5.3.5. *Fluorescence*

Fluorescence spectral signals are quite intense because of the efficient electronic excitation and fluorescent emission process. This enables measurements down to the nanosecond timescale on small samples with remarkable accuracy. The sensitivity is high over a wide concentration range. The spectral features are rather broad compared to those of other spectroscopic methods. Fluorescent techniques provide information on molecular arrangement and dynamics of micro-heterogeneous systems, and on micropolarity or microviscosity [321], [322].

Steady-state fluorescence spectroscopy and fluorescence lifetime measurements have been used for on-line reaction monitoring. Sources of fluorescence can be the reaction components themselves or a fluorescent probe that is added to enhance the selectivity of the measurement. Reaction monitoring with fiber-optic fluorescence sensors coated with fluorescent probe molecules is an emerging technology. In contrast, fluorescence signals in Raman spectroscopy are a source of disturbance because they originate from largely unknown impurities and mask the more specific Raman signal. The fluorescence technique is potentially much cheaper than Raman spectroscopy.

Fluorescence methods have been demonstrated to be useful in academic research for in situ monitoring of photoinitiated processes [323], curing [324-327] and polymerization [328-334]. Polymer colloids have been characterized by fluorescence quenching techniques in order to draw conclusions about the internal particle structure, transport phenomena, and particle flocculation of latices [335], [336]. A traditional field for the application of various fluorescence techniques are micellization behavior, molecular interactions, aggregation and clouding phenomena in surfactant solutions [337-340]. Optical fluorescence microscopy is capable of in situ chemical analysis of phase-separating polymer blends [341]. Various aspects of latex coalescence and film formation were studied by fluorescence techniques [342-345]. Various monitoring applications of fluorescence-based sensors and fluorescent probes are described in [346-351].

Fluorescence is not commonly used for on-line monitoring of chemical processes outside academia. Nevertheless, fuel in the cylinder wall oil film of a combustion engine has been detected in situ with a fiber-optic laser-induced fluorescence system [352]. Laboratory analytical work published in the past indicates that fluorescence methods may become especially useful to monitor biotechnological processes in the future.

5.3.6. *Other Spectroscopic Techniques*

In addition to the classical optical spectroscopic techniques discussed above laser-based spectroscopic methods are gaining increasing importance for process monitoring, gas analysis and trace element detection. Laser remote sensing techniques have been reviewed in [353].

In a laser gas sensor, the frequency of a tunable laser (diode) or a quantum cascade laser is adjusted to the absorption frequency of a specified gas component thus enabling the detection of molecules from the percent down to the ppm range in a rather complex matrix [354].

A high-energy laser pulse generates a plasma (a dielectric breakdown) on the material of interest in laser-induced breakdown spectroscopy, LIBS (or laser-induced plasma spectroscopy, LIPS). The resulting light emission can be analyzed with a spectrograph [355] to detect gaseous components in air [356], to determine the concentration and size of particles in water [357], to analyze steel components in the melt at elevated temperatures [358], and to identify plastics in a recycling stream [359].

In cavity ring-down spectroscopy (CRDS) a laser pulse is trapped in an optical cavity between two reflecting mirrors. The light intensity within the cavity decays as a function of time [360]. Absorbing species inside the cavity increase the total cavity losses thus enabling the detection of trace molecules in gases.

With the development of multivariate techniques, 3D spectral image analysis became very important for systems where spatial distribution of components is key to process monitoring, understanding, and control [361]. Applications include optical spectroscopic imaging from the UV to the IR wavelength range in absorption, fluorescence and Raman spectroscopy as well as magnetic resonance, X-ray fluorescence or XPS imaging.

[Top of Page]

6. Particle Size Analysis

An important characteristic of disperse systems (latices, emulsions, aerosols, suspensions, powders) is their particle size distribution. Particle sizes affect almost all properties of a disperse system (electrical, optical, rheological, etc.), as well as its stability. The evolution of the particle size distribution is a sensitive indicator of the progress of heterophase reactions, and it determines the application properties of final reaction products.

In this chapter, the main focus is on process particle size measurements in suspensions. For reviews of particle measurements, including in-process particle size distribution measurements and classification of microsized particles, see ⟶ Air – Measurement of Particles and ⟶ Particle Size Analysis and Characterization of a Classification Process.

6.1. Scattering Techniques

The scattering of electromagnetic waves (light, X-rays) and neutrons by particulate matter is a powerful tool for studying particle sizes and shapes, internal particle morphologies, particle dynamics (even under shear), the structure and dynamics of concentrated disperse systems, and particle charges. Wave interference yields information on particle sizes and shapes whenever the wavelength is of a comparable order of magnitude to the size of the scatterer.

Neutron and X-ray scattering both rely on the availability of corresponding neutron and radiation sources. Hence, their use for on-line size monitoring is extremely limited.

Particle size determination by light scattering techniques is a well established broad field of experimental methods. In addition to inelastically scattered light (Raman and Brillouin scattering), there are two basic approaches that exploit light scattering data for on-line particle size analysis: Static light scattering (SLS) and dynamic light scattering (DLS). SLS methods involve measurement of the time-averaged angular scattering as well as turbidity measurements (i.e., elastically scattered light). DLS measures the very small frequency shift of scattered light caused by translation and rotation of scatterers, or, in the FT time domain, fluctuations of the light intensity by correlation techniques. Thus, quasi-elastic light scattering (QELS), Rayleigh linewidth/spectroscopy, intensity correlation spectroscopy, and photon correlation spectroscopy (PCS) are synonymous with DLS.

Most of the light scattering techniques require extensive dilution of the extracted sample to avoid multiple-scattering effects. Automatic dilution devices are interfaced to the light scattering instrument for on-line applications. Direct optical access to the (undiluted) reaction mixture became possible with the development of fiber-optic probes and instruments.

For general reading see [362-364].

6.1.1. *Turbidimetry*

The turbidimetric experiment is rapid, precise, reproducible, and absolute (i.e., no calibration is required). Spectrophotometers are readily available. Turbidity gives a measure of the attenuation of light traversing a suspension or aerosol of nonabsorbing particles. Turbidimetry may require dilution of the mixture to avoid multiple scattering effects. The wavelength-dependent turbidity τ of a (diluted) sample provides information on particle size and concentration. Turbidimetric (and nephelometric) measurements at single wavelengths can be used to determine the concentration of particulate matter in liquid and gaseous process streams. Fraunhofer diffraction is used for particles sizes $\gg$ 1 µm.

Several publications describe the determination of particle size during emulsion polymerization by specific turbidity τ/φ or turbidity ratio measurements [365-367]. For a polydisperse suspension, the turbidity is related to the polymer volume fraction φ according to

$$\tau = \frac{3}{2}\varphi \frac{\int_0^\infty d_p^2 \cdot K\left(\frac{d_p}{\lambda_m}, \frac{n_p}{n_m}\right) f(d_p) \mathrm{d}d_p}{\int_0^\infty d_p^3 f(d_p) \mathrm{d}d_p} \tag{33}$$

where K, f, λ_m, and d_p are the scattering coefficient, the size distribution function, the wavelength of light in the medium, and the particle diameter, respectively. The quantities K and τ are functions of both the refractive indices of the particles n_p and of the medium n_m, and of the size of particles relative to the wavelength of the light d_p/λ_m. K can be calculated from the general Mie theory. If the size distribution function f is known, the particle size distribution can be estimated from specific turbidity measurements at several wavelengths. Despite the complicated dependence of τ on K and its direct dependence on the distributional form f (which is sometimes unknown), the specific turbidity can yield (1) the turbidity average particle diameter and the volume-surface average (i.e., $D_{3,2}$) diameter for small and large particles, respectively, for any value of $m = n_p/n_m$, (2) the weight average diameter for $m < 1.15$ and particles that are smaller than the wavelength of the light, and (3) an estimate of the weight average particle diameter in all other (monomodal) cases if a log-normal particle size distribution is assumed [366], [367].

Turbidimetry has been used to study the coagulation kinetics of aqueous dispersions [368], [369]. Efforts were made to extend the theoretical basis of turbidity to higher concentrations [370-372].

6.1.2. *Angular Static Light Scattering*

Angle-dependent static light scattering measurements with goniometers are not very suitable for on-line monitoring because of the time required and the demands discussed in ⟶ Plastics, Analysis – Light Scattering. A fast-response multichannel photometer capable of on-line monitoring even at moderate concentrations is described in [373]. The scattered light is simultaneously measured at 168 angles. This static light scattering instrument has a response time of 100 ms and an angular resolution of 1°.

Static light scattering at a fixed angle is valuable for studying the formation and aggregation of latex particles during emulsion polymerizations [374].

Several instruments are on the market which automatically dilute aqueous dispersions to a desired (extremely low) concentration and perform static multiple-angle light scattering. Different modes of light scattering (including Fraunhofer diffraction) are typically combined within one instrument to enable particle size analysis over a wide range of sizes.

6.1.3. *Dynamic Light Scattering*

Particles suspended in a continuous medium undergo random Brownian motion. Hence, the phase of light waves scattered by Brownian particles also fluctuates randomly in a time-resolved light scattering experiment. Their mutual interference leads to a net randomly fluctuating scattered light intensity $I(t)$ at the detector. The normalized autocorrelation function $g_2(t')$ of the light intensity as a function of time delay t' is given by

$$g_2(t') = \frac{\langle I(t) \cdot I(t+t') \rangle}{\langle I \rangle^2} \tag{34}$$

with $\langle I(t) \cdot I(t+t') \rangle$ ($t' \longrightarrow \infty$) = $\langle I \rangle^2$, the square of the average scattered light intensity. The field autocorrelation function $g_1(t')$ can be calculated if the Siegert relation holds

$$g_2(t') = A + B \cdot \left| g_1(t') \right|^2 \tag{35}$$

where A and B are instrument-related constants.

For a monodisperse system of Brownian particles, their diffusion coefficient D can simply be derived from $g_1(t')$ according to Equation 36

$$g_1(t') = \exp\left(-q^2 D t'\right) = \exp\left(-\Gamma t'\right) \tag{36}$$

where Γ is the decay constant. The scattering vector q is defined as

$$q = \frac{4\pi n}{\lambda} \sin\left(\frac{\theta}{2}\right) \tag{37}$$

where n, λ, and θ are the refractive index, the wavelength of light, and the scattering angle, respectively. The Stokes – Einstein equation relates the diffusion coefficient D to the hydrodynamic radius of the particles r_h

$$r_h = \frac{k_b T}{6\pi\eta D} \tag{38}$$

where k_b, T, and η are the Boltzmann constant, temperature, and viscosity of the dispersion medium, respectively.

For polydisperse systems, the autocorrelation function becomes a sum of exponentials with a distribution function $G(\Gamma)$

$$g_1(t') = \int_0^\infty G(\Gamma) \exp(-\Gamma t') d\Gamma \quad (39)$$

Particle size characteristics and distribution functions can be derived from measured autocorrelation functions by (1) directly inverting the Laplace integral equation (Eq. 39, a mathematically ill-conditioned problem), or (2) parameter fits, exponential sampling, regularization techniques, and histogram analysis.

Some advantages make dynamic light scattering attractive for on-line applications:

1. DLS is an absolute method; no calibration over time is necessary. With conventional DLS systems, accuracies of ± 1 % for the mean particle diameter can be achieved on a routine basis within a few minutes of measurement time. Care must be taken in reacting systems, where either temperature or viscosity change during the course of the reaction (see Eq. 38).
2. Particle properties such as density or refractive index do not affect the time behavior of the intensity fluctuations. Thus, the derived particle sizes are independent of the chemical composition of the particles.
3. In the single-scattering regime, diffusion coefficients and particles sizes are independent of particle concentration. This is an advantage over turbidity measurements, for which exact knowledge of the particle concentration is crucial for size calculations.
4. The use of inexpensive fiber-optics allows remote monitoring.

Two approaches are possible for on-line DLS measurements:

1. A device interfaced with the process line or reactor automatically captures a certain amount of the mixture and dilutes it sufficiently to avoid multiple scattering and allow noise-free measurements [375].
2. A fiber-optic probe (optode) is directly immersed into the mixture and both illuminates the colloidal system and collects the scattered light (FOQELS: fiber optic quasi elastic light scattering; FODLS: fiber optic dynamic light scattering [376-378]. This backscattering arrangement shortens the light path within the dispersion and makes the method applicable to turbid and concentrated systems. Monomode fibers further suppress multiple scattering effects [379].

Particle growth during emulsion polymerization has been studied by conventional DLS [380-384] and fiber-optic DLS [385], [386].

6.1.4. *Other Optical Techniques*

The effect of multiple scattering can be useful for studying highly concentrated dispersions. The path of multiply scattered light in the dispersion is similar to the path of particles or molecules under the influence of Brownian diffusion [387-390]. The concept of photon migration or diffusing wave spectroscopy (DWS) has already been applied to particle size determination.

Colloidal refractometry is an optical method that measures the refractive index *n* of a dispersion and analyzes it by using Mie theory. This method allows the structure of concentrated dispersions to be probed in the undiluted state. The measured values of *n* can provide a measure of the volume average particle size [391].

Frequency-domain photon migration has been employed to on-line monitor the variation of the active pharmaceutical ingredient concentration as well as changes of particle sizes in powders [392]. Chord length distributions of particles in the micron to millimeter range of emulsions and suspensions can be on-line determined by means of focal-beam reflectance measurement methods [393], [394]. Particle sizes ranging from 50 nm up to 3 μm are accessible by means of dynamic extinction spectroscopy [395]. Transmission fluctuation spectrometry [396], [397] combined with the autocorrelation technique covers a particle size range from 1 μm to 1 mm and a particle concentration of up to 12 % of flowing particle suspensions [398].

6.2. Separation Techniques

Recently, particle separation techniques (chromatography, fractionation) have been developed that are capable of on-line monitoring of particle size. These methods require fully automated removal of samples from the reaction mixture, dilution, and injection into the instrument.

Size exclusion chromatography (SEC) is applicable with porous (liquid exclusion chromatography LEC) and nonporous packing of the column (hydrodynamic chromatography HDC). Particles are separated because of the dependence of the rate of particle flow through gaps between packing beads on particle size. Larger particles are eluted first since they travel through larger gaps with higher flow rates. The main advantage of SEC methods is that particle size distributions can be obtained directly without any assumption regarding the mathematical form of the distribution. Calibration of the instruments with particle size standards is necessary but simple. Disadvantages are the limited resolution because of radial dispersion, and relatively long elution times of up to half an hour [399], [400].

Much higher resolution than with SEC methods can be achieved by capillary hydrodynamic fractionation (CHDF) [401] and field-flow fractionation (FFF). The fluid in a CHDF capillary has a parabolic velocity profile with the greatest fluid velocity at the center of the tube and zero velocity at the wall. Particles move radially due to their Brownian motion. Larger particles are unable to approach the wall as closely as smaller particles. Hence, larger particles travel through the tube faster than smaller particles. This separation effect is exclusively a function of particle size; it is independent of particle density. The efficiency of capillaries for separating particles depends on the eluent viscosity, the flow rate, and the capillary diameter. The optimum particle size for CHDF analysis is < 1 μm. The main disadvantage of this method is the possibility of capillary

clogging. The CHDF method has been used to monitor the evolution of the particle size distribution during emulsion [402] and miniemulsion polymerization [403].

Field-flow fractionation (FFF) is a family of chromatography-like elution techniques based on influencing the rate of particle flow through a narrow channel by applying an external field perpendicular to the flow direction [404-408]. The external field separates particles of different sizes by driving them into different localized laminas (< 10 µm thick) of the parabolic velocity profile of the eluent within the fractionation channel thus causing their separation. In contrast to CHDF, FFF does not necessarily require calibration of the instrument. Retention times for particles of different sizes can be calculated directly from first principles.

FFF can be classified according to the type of applied field into sedimentation, thermal, electrical, cross-flow, and steric. The most suitable external fields for particle size analysis are cross-flow and sedimentation fields. The particle separation in cross-flow occurs according to the hydrodynamic radius of the particles, whereas their effective weight separates particles in a sedimentation field. The particle size ranges for cross-flow and sedimentation field FFF are 10 nm to 100 µm and 50 nm to 100 µm, respectively. Recently, FFF (including thermal FFF) has been used to characterize the size and composition of core/shell latices [409]. On-line coupling of FFF with multi-angle laser light scattering is described in [410], [411].

[Top of Page]

7. Chromatography

7.1. Introduction

Chromatographic methods (especially gas chromatography) are the most widely applied family of techniques for determining the chemical composition of reaction mixtures. For a discussion of basic principles, techniques, instrumentation, sample extraction and preparation, applications (including on-line applications), advantages and limitations see ⟶ Basic Principles of Chromatography, ⟶ Gas Chromatography, and ⟶ Liquid Chromatography.

A comprehensive discussion of (gas) chromatography for on-line monitoring of chemical reactions has been published [412]. Process applications can also be found in [413-416].

The term “process chromatography” almost exclusively refers to the application of gas chromatography (GC) to monitor chemical reactions or process streams. Advantages of GC are its simplicity, high sensitivity and selectivity, and ease of automation. Liquid chromatography (LC, including HPLC), supercritical fluid chromatography (SFC), electrophoresis (EP), and ion chromatography (IC) are less common in process monitoring because of their higher maintenance requirements, greater complexity, and higher costs for purchase and maintenance.

Because GC (like other chromatographic methods) is an inherently discontinuous *batchlike* analytical technique which is applied to *continuously* monitor a chemical system in a nonbenign environment, there are distinct differences between laboratory and process instruments, although the basic components of a gas chromatograph can be found in instruments of both categories (Fig. 19) and the basic principles of separation are the same.

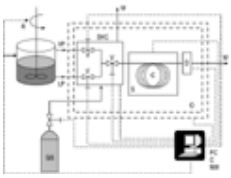

Figure 19. Schematic of a process GC analyzer GS: gas supply for carrier, auxiliary, and purge gases; V: valves for preseparation, column switching, and injection; VP/LP: vapor-phase/liquid-phase sampling; SHC: sample handling and conditioning system; R: reactor; S: separation unit; C: column(s); D: detector; O: thermostatically controlled oven; PC/C/NW: personal computer/controller/network; W: waste gases

In contrast to laboratory instruments, process analyzers have to meet stringent safety and explosion-proof requirements because of their close vicinity to the product line. The typical GC process instrument is designed to detect and monitor only one or a few (previously known) components of the reaction mixture so as to control a chemical reaction or process stream. This requires special injection valves and column-switching devices. As in conventional GC, the amount of sample required for process GC is extremely small. Hence, the sampling system must ensure that the sample is representative of the process to be monitored. Often, a sample conditioning system is necessary to further process the sample, which may still be reactive, for GC analysis.

The nonspecific character of detection based exclusively on retention time (i.e., no positive identification) makes GC preferably suitable for process monitoring of continuous processes with small variations in composition. Different molecules may have the same retention time, confusing quantification of the chromatograms and leading to misinterpretation of overlapping peaks. Additionally, GC analysis requires time for separation and detection. This time interval must be reasonably short compared to both the run time of the reaction and the time required to feed the analytical result into a process controller, and to influence the reaction.

7.2. GC Hardware Components

Carrier Gas. In principle, the same carrier and auxiliary gases can be used in both laboratory and process GC analyzers (usually hydrogen, helium, argon, synthetic air, or nitrogen, depending on the application). Most high-purity gases are supplied in containers (e.g., high-pressure gas cylinders). For an on-line operation of the GC analyzer, the gas storage and delivery systems should be equipped with valves that switch automatically between empty and full containers.

Sampling. The quality of a GC analysis of a reacting mixture is only as good as the quality of the samples withdrawn from it. Because of the very small amounts of sample necessary for GC, it is crucial to withdraw samples that are representative

of the mixture. The sampling point (sample probe) should be located at a point in the bulk of the medium where it is homogeneous and well mixed so as to obtain the “true” bulk composition of the mixture. Close to the walls, adsorption and deposition phenomena and stagnant flow may locally change the overall composition.

From an analytical point of view, the best location for the analyzer would be right next to the process line or sampling loop. However, a number of factors and restrictions have to be taken into consideration for placing the chromatograph relative to the sampling point, e.g., 1) space available for installation, 2) maintenance requirements, 3) safety requirements, 4) accessability of sampling point and chromatograph, 5) connection to utilities and supplies (carrier gas, electricity, compressed air), 6) optimization of instrument lag time versus control cycle time, 7) minimization of sample transport, 8) reactivity/stability of the analyte. In most applications, a sample transfer line bridges the distance between sampling point and analyzer. Length and diameter of the transfer line must be optimized for (high) analyte flow. Pressure drop and temperature changes along the line must be controlled. When the analysis is complete, the sample is returned to the process by a separate line or disposed of along with the carrier/purge gas.

In many processes, GC samples are taken at elevated temperature and/or pressure. These samples may contain unconverted components, other reactive ingredients, water vapor, liquids, or solid impurities. Hence, process samples must be appropriately conditioned before feeding them into the GC analyzer. The sample conditioning system performs depressurizing and temperature reduction as well as filtration and removal of liquids (e.g., by vaporization). Its different elements may be close to the sample point (pressure reducer, filter) or close to the analyzer (temperature control, flow control). The conditioning system ensures that the temperature of gaseous samples is held above the dew point to avoid condensation and that the temperature of liquid samples is well below the boiling point to avoid bubble formation in the line.

Sample transfer and conditioning are the most time-consuming part of an on-line GC analysis. The lag time between process and analysis is on the order of a few minutes and must be taken into account in designing reaction control systems based on conventional GC. With microsensor GC systems, timescales on the order of (milli) seconds for on-line analysis are possible.

Analyzer Oven. In contrast to most laboratory instruments, the actual process chromatograph (i.e., injection system and valves, column(s), detector(s), controls) is placed in a single temperature-controlled oven. Temperature control of better than ± 0.05 K suppresses fluctuations in temperature-sensitive processes such as gas flow, mobile-phase diffusion, adsorption/desorption, and chemical interactions. Since detection is solely based on retention time, all processes that can influence retention times must proceed unaltered to avoid loss of calibration under automated operating conditions. Constant temperature is usually guaranteed by circulating heated air within the process analyzer oven or by using high-mass metal heat reservoirs. In some applications, tighter temperature control can be circumvented by an independent delayed injection of a standard or by adding an internal standard to the carrier gas.

Pressure control and programming is dominant in process chromatographs, although some temperature programmable analyzers and gradient systems have been recently described in the literature.

Valves and Columns. In contrast to laboratory instruments, a process chromatograph is usually designed to monitor only one or just a few components of the reacting system for control purposes. Exceptions include some petrochemical applications, in which up to several hundreds of components are monitored simultaneously. Additionally, the process GC analyzer must operate unattended in a rather harsh environment over extended periods of time, but still provide reproducible and stable data. Of special importance is a constant sample volume/mass for injection (and volume of standard, if necessary). Therefore, dedicated valve systems for sample injection and column switching have been developed [412] (see ⟶ Liquid Chromatography – Sample Valve Injectors, ⟶ Liquid Chromatography – Coupling Techniques). For gaseous samples, the sample size is normally defined by the volume of an external sampling loop connected to different valve ports. The injection volume of liquid samples can be defined by the size of internal grooves within the valve or by external sampling loops.

Sliding, rotary, and diaphragm valves with a wide variety of internal pathways and port configurations are commonly used in process chromatography of gaseous samples. For liquid samples, piston (syringe) valves may also be applied. Valves are heated externally or internally. Depending on the component(s) of interest within the complex reaction mixture, cutting (mainly heart-cut technique) and other column-switching techniques (e.g., backflushing) or multicolumn analysis are often combined with precolumn separation (see ⟶ Liquid Chromatography – Column Switching). The main reason for using column-switching techniques in on-line process GC is to drastically reduce the total analysis time and to fully separate the components of interest from the mixture. In general, the same types of valves are used for injection and column switching.

Traditionally, the use of packed columns predominates in process GC. As in laboratory GC, packed columns are being replaced by capillaries in an increasing number of on-line applications. With capillary columns, chromatographic separation is improved, resolution is higher, and analysis time is shorter than with packed columns because of the much simpler flow and diffusion pattern of the analyte. However, capillaries are normally less sensitive, more expensive, and not so easy to use as packed columns. A major drawback of capillary columns in process applications is their sensitivity to flow, pressure, and especially temperature fluctuations.

Columns, their supports, coatings, and packing materials have to withstand possible corrosive action of ingredients of the reaction gas mixture and should be inert to reactions with solute molecules (e.g., by silylation of the column packing material or capillary coatings). The presence of water vapor and oxygen traces in process gases may also be of relevance to column selection.

Detectors in process chromatographs have to comply with the overall safety, robustness, and reliability standards of an on-line instrument. The overwhelming majority of GC detectors utilized in process analyzers are of the thermal conductivity (TCD) and flame ionization (FID) types because of their ease of use and sensitivity to the process components of interest. TCDs are sensitive to changes in the bulk thermal conductivity of the carrier gas in the presence of the analyte, which

gives a concentration-dependent signal, whereas FIDs sense the number of certain ionized molecules entering the flame with a specific flow rate. In some rare cases, photoionization (PID) and flame photometric (FPD) detectors are applied.

Several methods have been utilized in process GC to convert detector output signals (peak areas) into desired analyte concentration values, such as the separate injection of calibration standards, the use of internal standards, and calculations based on relative response for the different detector types.

The basic operations of a GC can also be miniaturized and micromachined on a silicon wafer. Micro-GC devices are compact, more robust, and faster than classical GCs what makes them very attractive for process monitoring applications [417].

7.3. Applications of Gas Chromatography

GC has a long tradition of use for controlling processes and process streams in industry. On-line GC is one of the most important techniques for monitoring distillation, rectification, and conversion processes in refineries and chemical and pharmaceutical plants. It is extensively applied in catalysis research, development, and production to evaluate catalysts and study reaction mechanisms and kinetics. Another field of use is in metallurgical smelters for gas and sulfur analysis in metals. It is widely utilized as a part of workplace safety systems, e.g., to ensure air quality in production units, pilot plants, and laboratories.

On-line GC has proved useful for determining organic substances in water [418], [419], wastewater analysis [420], and air monitoring [421-423]. Reports have been published on the on-line analysis of petroleum [424], [425] and gasoline [426-428], the determination of octane numbers with *n*-alkanes as reference [429], the on-line monitoring of methyl formate conversion to methanol and carbon monoxide [430], Fischer – Tropsch synthesis [431] and thermolysis/decomposition of organic compounds [432], [433]. The application of gas chromatography (among other techniques) for on-line monitoring of polymerization reactions is reviewed in [434]. Gas chromatography of the reaction mixture [435], [436] and the head space of the reactor [437], [438] can be applied for on-line determination of residual monomer composition in emulsion polymerization.

Head-space GC in liquid reaction mixtures is not straightforward for direct reaction monitoring because it requires knowledge of the partitioning of each component between the gas phase above the reaction mixture and the liquid mixture, and, especially in multiphase reactions, between the different phases of the reaction mixture (e.g., aqueous phase, monomer droplets, and polymer particles in an emulsion polymerization). Additionally, the head space should be in thermodynamical equilibrium with the fluid reaction mixture.

The utilization of (mass) sensors for gas and liquid analysis has been reviewed in [439-441], In-line applications of chemical sensors to monitor polymerization reactions has not been described in the literature so far. Such sensors should have a very high selectivity so as to detect components well below the percent range in a matrix of solvent or solvent vapor (including water) containing a large number of ingredients, some of them with a chemical structure comparable to that of the analyte. Additionally, the sensors would have to deal with precipitates on their surfaces.

7.4. Other Chromatographic Techniques

Applications of liquid chromatographic systems to monitoring chemical and biochemical processes are described in [442-448]. SAW sensors are useful gas sensing tools for on-line monitoring of the mobile phase of process HPLC with changing mobile phase composition or gradients [449]. On-line ion chromatography has been used to monitor adipic acid production [450], trace metal ions [451] and cations [452], and oxides in flue gases [453]. Reviews on process ion chromatography can be found in [454] and [455]. In some rare cases, capillary electrophoresis [456], [457] and supercritical fluid chromatography [458], [459] have been described as on-line reaction monitoring techniques.

[Top of Page]

8. Electroanalytical Methods

8.1. Introduction

Electroanalytical methods are based on the interaction of electrical fields with chemical processes or on chemically induced electrical signals on electrodes. The major advantage of electroanalytical methods is the electrical output of the measurement (i.e., electrical current, potential, resistance, or charge), which can directly be processed electronically without any further transformation, in contrast to other monitoring techniques. Although the main application for electroanalytical techniques is process monitoring in aqueous solutions, they can also be applied to nonaqueous and nonliquid systems.

In most electroanalytical measurement systems, the sensing electrodes are in direct contact with the reaction medium and all of its components, not only with the species of interest. This can lead to cross-sensitivity and poisoning, and to abrasion, deactivation, and contamination of the active electrode surfaces. Additionally, the analytical information of interest must be extracted from the integral signal. Calibration of the sensing system is typically necessary at the beginning of a measurement cycle and subsequently in regular intervals. Accumulation of static electricity within the sensing system and on its body must be avoided.

Compact titration configurations for on-line process analytics are commercially available not only for electroanalytical applications such as acid – base titrations but also for precipitation and complexometric titrations [460].

8.2. Conductometry

Conductometric measurements rely on the existence of charges in the medium that are mobile in an applied electrical field (e.g., ions in electrolyte solutions or melts, electrons in metals). In conductometry, no reaction takes place at the

electrodes. Electrical conductance of the medium or electrical capacitance of a measurement cell is the measured quantity. Conductometry in solutions and melts is a nonspecific method because all anions and cations contribute to the medium's conductivity. However, the method is very robust, precise, sensitive, fast, and does not require extensive maintenance.

Concentration c_i and μ_i of all the ions *i* present in a medium determine its conductivity Λ (and conductance *G*)

$$\Lambda = kG = kF\sum_i c_i n_i \mu_i \tag{40}$$

where n_i is the ion valency, *F* is Faraday's constant, and *k* a cell constant. This linear relationship between concentration and conductivity is the basis for concentration measurements in the lower concentration range ($\leq 0.1\ \Omega^{-1}\ cm^{-1}$). For higher electrolyte concentrations, Equation 40 becomes nonlinear because of ionic interactions. Additionally, polarization effects may occur at the electrodes. The temperature dependence of the ionic mobility μ leads to an almost (quasi) linear dependence of Λ on temperature in the temperature range common for conductivity measurements in electrolyte solutions (0 – 100 °C).

Two different principles of conductivity measurements are in industrial use: 1) two- or four-electrode conductive cells with direct electrode contact to the medium, and 2) inductive, electrodeless, and contactless devices.

Contact electrode cells are applied in nonaggressive media. The cell itself is typically part of a Wheatstone bridge operating with low-frequency alternating current to avoid electrode polarization. Two-electrode cells are in use in the conductance range 10^{-8} to $10^{-1}\ \Omega^{-1}\ cm^{-1}$, whereas four-electrode cells are suitable for the higher conductance range of 10^{-3} to $10\ \Omega^{-1}\ cm^{-1}$.

Electrodeless inductive measurements are the method of choice for highly conductive systems; for aggressive, abrasive, or corrosive media; and for high-temperature applications. The medium to be measured is circulated through a loop; two transformer coils are attached to the inert wall material of the loop pipe (Fig. 20). A generator supplies alternating voltage to the primary coil. The induced alternating voltage at the secondary coil depends on the conductivity of the solution circulating through the loop. For contactless measurements, the medium's conductance is typically much higher than $10^{-1}\ \Omega^{-1}\ cm^{-1}$. Inductive devices are less sensitive than electrode cells.

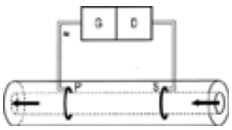

Figure 20. Schematic of an electrodeless conductivity measurement device G: generator; D: detector; P: primary coil; S: secondary coil; ≈: alternating current

The electrical conductivity Λ_d of a disperse system is a function of both the volume fraction of the disperse phase φ ($\varphi \ll 1$) and the electrical double layer properties of the particles (surface conductivity Λ_s, Zeta potential ζ)

$$\Lambda_d = \frac{1}{f}\Lambda_0 + \Lambda_S \tag{41}$$

where Λ_0 is the conductivity of the dispersion medium and *f* a form factor accounting for the shape and the volume of the (nonconducting) particles [461]. Thus, conductivity measurements provide information on particle formation, disappearance of particles, phase inversions, and particle swelling in heterophase reactions such as crystallization and emulsion and suspension polymerization [462-466].

Electrical resistance tomography is a rather new technique for on-line monitoring of chemical reactions of particulate systems in tank reactors. The voltage between individual electrodes of a system of electrodes distributed over one or several layers in the reactor wall is used to calculate an image of the electrical resistance distribution within the reactor which is indicative of changes of the dispersion state of multiphase systems [467].

Most applications of conductometry deal with monitoring of process streams rather than direct reaction monitoring. Typical application fields are drinking water purification and tap water monitoring, wastewater treatment, cleanness of ultrapure water in the electronics industry, distillation processes, neutralization and precipitation reactions, concentration monitoring of acids (e.g., in production of concentrated sulfuric acid), lyes, and salt solutions, and monitoring of process streams in dairies and breweries.

Electrostatic sensors can be used to on-line measure the mass median size of particles in a dilute-phase pneumatic suspension. The sensor configuration includes a wire-mesh electrode, signal conditioning electronics, and digital signal processing algorithms [468].

8.3. Potentiometry

Potentiometric techniques are characterized by the fact that electrode reactions at equilibrium are involved in the potential measurement. In potentiometry, the actual electrical current through the electrode is zero. The electrode potential cannot be measured directly but only as a difference with respect to the potential of another electrode (reference electrode). Electrode potentials and the potential difference between two electrodes can be calculated according to Nernst's equation (Eq. 42)

$$U_{\text{electrode}} \sim \frac{RT}{F} \ln \left(\frac{a_i}{a_{i,\text{ref}}} \right) \quad (42)$$

where R is the gas and F is Faraday's constant. The ion activities a_i can be activities of two different ions within the same solution (redox pair), activities measured with two different electrodes, or activities at two different concentrations or partial pressures.

Ion-Selective Electrodes. The swelling glass layer of ion-selective electrodes is designed to be sensitive to a certain type of ions and insensitive to others. This can be achieved by using membranes containing inorganic salts of the ions of interest (e.g., Na^+, K^+, Cu^{2+}, NH^{4+}, Cl^-, F^-, I^-). Conditioning of the mixture to a certain conductivity and pH level is a prerequisite for ensuring precise concentration measurements with ion-selective electrodes. Hence, the use of this type of electrodes for direct (in-line) insertion into a reaction mixture is limited [469], [470].

pH measurement is the most important and most widely used potentiometric technique in aqueous solutions or water-containing systems. Measurements of pH are crucial for monitoring processes in the chemical and pharmaceutical industries, the food and beverages industry, waste-treatment plants, environmental protection, and biochemical operations [471]. In particulate systems, the pH has a dramatic influence on colloidal stability (e.g., cosmetic products, emulsion and suspension polymerization). Care must be taken to take into account temperature variations of the reacting system because of the strong temperature dependence of electrode potentials.

pH measurements are easily performed with standard glass electrodes. For process applications, the most common pH-sensitive membranes consist of lithium glasses. Other alkali metal glasses are prone to increased hydrolysis and instability, especially in alkaline solution. Several types of glasses are available for different process environments (standard, highly alkaline, highly acidic, high temperature). Problems may arise with glass electrodes due to contamination of the surface by precipitation or film formation. Cleaning and maintenance of glass electrodes on a regular basis are critical for successful on-line pH measurements.

Glass electrodes with counterpressure are applicable for on-line monitoring of reactions up to temperatures of about 80 °C and pressures up to 60 bar. Mechanical instability is a major drawback of glass pH electrodes in high-pressure applications.

Solid-state ion-sensitive field effect transistor (ISFET) sensors show faster response, lower alkaline errors, and improved mechanical stability than glass electrodes. However, they also suffer from surface contamination. The applicability of ISFET pH sensors is limited to pressures well below 50 bar and temperatures up to 85 °C [472]. There is no solid-state sensor on the market applicable for pressures above 60 bar. An as-yet not commercially available ZrO_2 sensor for pH measurements under high pressure (83 bar) and high temperature (285 °C) is described in [473]. A number of different optical and fiber-optical pH sensors have been developed [474-477].

Redox Potential. The oxidizing or reducing character of a solution can be measured with (inert) precious metal electrodes (Pt, Au) relative to a reference electrode. Redox potential measurements are highly nonspecific because all redox couples of a mixture contribute to the potential. The redox potential also strongly depends on pH. Thus, the applicability of redox potential measurements is mainly limited to determining qualitatively, e.g., nitrites in wastewater (denitrification), cyanides, copper, and chromium in galvanic wastes, or chlorine in public swimming pools.

Potentiometric Oxygen Analysis. Solid-state electrolytes that exhibit ionic conductivity at elevated temperatures (> 500 °C) can also be used for potentiometric measurements. Porous ZrO_2 membranes covered with platinum grids have been used to determine oxygen in gases potentiometrically. Different oxygen concentrations on the two sides lead to a potential difference across the membrane in proportion to the difference in oxygen partial pressure.

8.4. Amperometry, Voltammetry, and Coulometry

Amperometric, voltammetric, and coulometric techniques are not so widely used as other electroanalytical methods for reaction monitoring. They are characterized by electrode reactions at polarized electrodes that are detected by current – voltage measurements. Because of the electrode reactions the electrodes have to be cleaned on a regular basis, and a certain fraction of the mixture under investigation is consumed during the measurement.

Applications include the amperometric determination of chlorine, dissolved oxygen, hydrazine, and ozone in pulp-bleaching baths; tap, industrial, and wastewater, amperometric monitoring of hydrogen peroxide in textile bleaching baths [478], the amperometric monitoring of gases and air at work places, the voltammetric determination of heavy metals in waste streams, and the coulometric analysis of sulfur compounds.

[Top of Page]

9. Miscellaneous Methods

9.1. Mass Spectrometry

In the 1990s, mass spectrometry (MS) emerged as technique for on-line analytical purposes. The method is highly sensitive and fast with an extraordinary molecular selectivity. The detection principle of MS is based on an ionization of chemical species followed by a separation of the fragments according to their mass-to-charge ratio (see ⟶ Mass Spectrometry). Due to the wide variety of ionization techniques, most types of sample can be analyzed by MS. Ionization techniques for process applications include electron impact (EI), resonance-enhanced multiphoton ionization (REMPI), microwave plasma ionization, thermal ionization, and electrospray ionization (ES). Direct capillary and membrane inlets are most commonly used for sample introduction. MS process applications are reviewed in [479-483].

In an increasing number of process applications MS is interfaced to pre-separation techniques such as GC or LC to improve resolution and selectivity in the analysis of complex (reaction) mixtures. Reduced costs, size, and complexity of the instruments have even made portable MS devices possible. Efforts have been undertaken within the last couple of years to miniaturize MS instruments ("MS on a chip").

The traditional on-line fields of use of MS are process gas stream and environmental gas analysis, environmental and wastewater monitoring, and monitoring of fermentation processes. Trace amounts of aromatics and chlorinated compounds in flue gases of industrial incinerators have been detected by REMPI-MS [484-486].

MS has also been applied to monitoring liquid-phase and gas-phase chemical reactions. On-line MS is useful for monitoring alkanes and alkenes in rubber production [487], monomers and solvents in polyethylene production [488], monitoring the photocatalytic degradation of phenol and trichloroethylene [489] or electrochemical polymerization of aniline [490].

The modern semiconductor industry relies heavily on MS process monitoring [491], [492]. MS can be used to monitor gaseous impurities in high-purity gases, to study plasma etching processes [493], molecular beam epitaxy, and chemical vapor deposition processes [494].

9.2. Densimetry and Dilatometry

The physical basis of densimetry and dilatometry is the mass density difference (if any) between initial reaction mixture and the final reaction product. As the reaction proceeds, this density difference leads to an increase or decrease in the density of the mixture $\varrho(t)$ and to an overall shrinkage or expansion, respectively, of the volume of the reaction mixture V. Density changes can be measured continuously with responses on the order of one measurement per second, while the repetition time for measurements of volume changes is in the millisecond range. Therefore, the extent of reaction can be calculated at any time if the amount of ingredients initially charged to the reactor and/or continuously fed to the reaction mixture is known.

The following factors must be considered in applying these techniques:

1. For a quantitative analysis of density changes, it has to be checked carefully whether the reaction mixture behaves as an ideal mixture or not; that is, does Equation 43

$$V_{\text{total}} = \sum_i V_i + V_{\text{excess}} \tag{43}$$

   with $V_{\text{excess}} = 0$ hold, where V_{total}, V_i, and V_{excess} are the volume of the reaction mixture, volume of component i, and excess volume, respectively.
2. Accurate densimetric and dilatometric measurements rely substantially on the accuracy of temperature control to be better than ± 0.1 K (in some cases even better than ± 0.01 K).
3. As with all other "integral" methods, a single variable (density, volume, temperature, pressure) is insufficient to describe multi-component reaction mixtures.

Density measurements are suitable for reaction monitoring of polymerizations (in bulk, solution, emulsion, suspension) because of the considerable density differences between monomers and polymers (Table 6). In a two-component system, the extent of reaction (e.g., conversion $x_r(t)$ in polymerizations) can simply be calculated according to Equation 44 by assuming ideal mixing behavior

$$X_r(t) = \frac{\frac{1}{\varrho_{\text{start}}} - \frac{1}{\varrho(t)}}{\frac{1}{\varrho_{\text{start}}} - \frac{1}{\varrho_{\text{end}}}} \tag{44}$$

where ϱ_{start} and ϱ_{end} are the density of the mixture at the beginning and end of the reaction.

Table 6. Densities of monomers and polymers at 20 °C

Monomer	Density, g/cm^3	
	Monomer	Polymer
Acrylic acid	1.051	1.37
Acrylonitrile	0.806	1.17
1,3-Butadiene	0.621	0.97
Butyl acrylate	0.899	1.08
n-Butyl methacrylate	0.886	1.06
Ethyl acrylate	0.924	1.12
2-Ethylhexyl acrylate	0.887	0.99
Ethyl methacrylate	0.914	1.13
Methyl acrylate	0.954	1.22
Methyl methacrylate	0.944	1.19

o-Methylstyrene	0.916	1.01
Styrene	0.906	1.04
Vinyl acetate	0.932	1.18
Vinyl chloride	0.911	1.39
Vinyl methyl ether	0.750	1.06

Most of the common principles to perform density measurements (pyknometer, buoyancy) are unsuitable for reaction monitoring. Methods that have been applied to process monitoring include hydrostatic pressure measurement, the maximum bubble pressure technique, oscillations of U-shaped tubes, cylinders or plates, and absorption of radioactive radiation or X-rays.

A rather robust technique is the measurement of the resonance frequency of a U-shaped tube (Fig. 21). This resonance frequency is a very sensitive function of the mass of fluid within the tube, and, hence, of its density. The densimeter has to be calibrated with a liquid of known density ϱ_{cal} that gives a period $T_{osc,cal}$ of the U-tube oscillations. The density of the mixture $\varrho(t)$ can be derived from the change of the period of oscillations $T_{osc}(t)$ provided the temperature T of the mixture is held constant (Eq. 45)

$$\varrho(t) - \varrho_{cal} = \text{const.}\left(T^2_{osc}(t) - T^2_{osc,cal}\right) \tag{45}$$

Figure 21. U-shaped tube for density measurements (courtesy of Anton Paar KG)

To obtain noise-free results, gas bubbles should be avoided while circulating the reaction mixture through the tube densimeter. In low-viscosity media, bubble traps can prevent bubbles entering the densimeter tube. A prerequisite for the application of Equation 45 is that the viscosity of the calibration liquid is roughly equal to the viscosity of the mixture because of its influence on the tube attenuation. The accuracy of density measurements with commercially available oscillating tubes can reach the range 10^{-6} to 10^{-7} g/cm^3. Applications of U-shaped tube density measurements to monitoring reactions can be found in [495-497].

In addition to gas bubbles and temperature uncertainties, film formation on the inner surface of the tube, formation of sediments, and clogging of the pipe may cause problems with this method. These problems can be partially overcome by the inverse arrangement: a tuning fork inserted into the mixture. Again, the change of resonance frequency of the fork is related to the density of the surrounding liquid, and the amplitude with which the fork oscillates (i.e., the attenuation) is a function of the viscosity of the medium.

The maximum bubble pressure technique is based on the pressure p_{bubble} of a spherical gas bubble of radius r at a capillary tip immersed in the reaction mixture at depth h (Fig. 22). This pressure is the sum of the hydrostatic pressure of the liquid above the capillary tip, and the pressure generated by the liquid – bubble interfacial tension γ

$$p_{bubble} = \varrho g h + \frac{2\gamma}{r} \tag{46}$$

where g is the acceleration due to gravity [498], [499]. Thus, p_{bubble} depends on both the immersion depth of the capillary tip into the liquid, and the capillary and, therefore, bubble diameter.

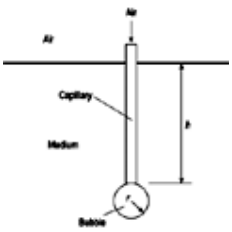

Figure 22. Schematic of maximum bubble pressure method

For density measurements, two capillaries of equal diameter are inserted into the mixture at different depths h_i beneath the surface. This leads to a difference of pressure readings Δp_{bubble} of the two capillaries, and

$$\varrho = \frac{\Delta p_{bubble}}{(h_1 - h_2)g} \tag{47}$$

The internal bubble pressure due to surface tension is identical for the two capillaries.

In addition, the two-capillary principle can also be used to measure the surface tension of the liquid. In this case, two capillaries of different diameters r_1 and r_2 are immersed into the liquid at the same depth and experience the same hydrostatic pressure but different internal pressures due to different bubble curvatures. Thus,

$$\gamma = \frac{\Delta p_{\text{bubble}}}{2\left(\frac{1}{r_1} - \frac{1}{r_2}\right)} \tag{48}$$

Corrections must be made to account for the influence of capillary forces and nonspherical bubble shapes [498]. In stirred or turbulent media, bubble shear-off may cause problems as well as coagulum formation or clogging of capillaries in film-forming media. The accuracy of bubble densimeters is on the order of 10^{-3} g/cm^3.

The hydrostatic pressure p_h of a liquid measured at a depth h is

$$p_h = \varrho g h \tag{49}$$

The hydrostatic pressure difference Δp_h between two pressure sensors at different depths of distance $\Delta h = h_1 - h_2$ is directly proportional to the liquid density

$$\varrho = \frac{\Delta p_h}{\Delta h g} \tag{50}$$

The physical basis of Equations (49) and (50) is essentially identical. In stirred or flowing media, the hydrodynamics may add considerable noise to the pressure readings of the sensor.

The absorption of radioactive radiation (including X-rays) depends on the molecular/atomic absorption coefficient, the density (concentration) of the ingredients, and the thickness of the material through which the radiation passes (path length d). After calibration of the absorption of both the source/cell/detector arrangement and the mixture, the density can be calculated according to

$$I_d \propto I_0 \exp(-\varrho d) \tag{51}$$

where I_0 and I_d are the intensity of the radiation source and the detected radiation, respectively.

Radiometric density determination is advantageous in systems under very high pressure and in corrosive media because neither the source nor the detector is in direct contact with the mixture.

The application of dilatometry to reaction monitoring is discussed in detail in [500]. Time-resolved measurements of volume changes for transparent and nontransparent samples with a laser scanning dilatometer show a simple linear relation between shrinkage and chemical conversion during epoxy curing [501]. In general, dilatometry is a very useful tool for academic research but only of limited value for reaction monitoring in (continuously stirred) industrial reactors.

9.3. Rheometry

The theory of rheology, instrumentation and applications are extensively described in [502-504] (see also ⟶ Rheometry, ⟶ Fluid Mechanics).

Various processes in the plastics industry are accompanied by viscosity changes of the reaction mixture:

1. Most of plastics manufacturing processes lead to a substantial increase of viscosity because of growing chains, branching, and cross-linking. Rheological characterization of plastics is essential for understanding and influencing their processability.
2. Curing reactions transform liquids into gels and solids. Thus, the resistance of the material to flow changes by orders of magnitude.
3. Melt flow index is an important parameter for monitoring and controlling reaction extrusion, reaction blending, and reaction injection molding [69], [70].

Increasing viscosity in noncross-linked systems can be related to the molecular mass and concentration of growing chains. For linear polymers, the bulk viscosity η is a function of the weight average molecular mass M_w similar to the Kuhn – Mark – Houwink equation for the intrinsic viscosity [η] of polymers in solution

$$\eta = K_\eta M_w^{a_\eta} \tag{52}$$

where K_η is a parameter that depends on interchain friction. The exponent a_η is about 1 – 2 for short chains, and increases to 3.4 above a critical molecular weight $M_{w,crit}$ [505]. Equation 52 holds also for branched polymers if M_w is replaced by gM_w, where the contraction factor g is defined as the ratio of the radius of gyration of the branched molecule to that of a linear molecule of the same weight average molecular mass.

In cross-linking reactions, the viscosity of the mixture increases dramatically while approaching the gel point [506], [507]. After reaching the gel point, the cross-linked reaction mixture behaves like a soft solid. For temperatures above the glass transition temperature T_g the mechanical properties of the material may be treated by rubber elasticity theory.

The viscosity of disperse systems depends on the volume fraction φ of disperse material, particle shape and size distribution, interparticular interaction, etc. For monodisperse spherical particles at low φ, viscosity is proportional to the volume fraction of the disperse phase

$$\eta = \eta_0(1 + 2.5\varphi) \tag{53}$$

where η_0 is the viscosity of the dispersion medium. For higher volume fractions and systems with dominant interparticle interaction, numerous theoretically based, semi-empirical, and empirical $\eta - \varphi$ relations can be found in the literature [508], [509]. Thus, viscosity reflects changes of particle number and size [510] in heterophase reactions (e.g. emulsion formation, emulsion polymerization, suspension polymerization, dispersion polymerization).

Principles and methods for on-line rheometry include rotational and capillary process viscometers (see ⟶ Rheometry), as well as the application of ultrasound (see Chap. Ultrasonic Methods), and tube or fork oscillations (see Section Densimetry and Dilatometry).

Widely used techniques for measuring melt viscosity in extruders are capillary viscometry and pressure drop analysis [69], [70], [511], [512]. The pressure drop Δp of melt flow through a nozzle is directly proportional to the melt viscosity according to the Hagen – Poiseuille law

$$\Delta p = -\frac{8\eta\bar{v}l}{R^2} \tag{54}$$

where $\bar{v}$, l, and R are the average velocity in the pipe, its length, and diameter, respectively.

9.4. NMR Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy (see ⟶ Nuclear Magnetic Resonance and Electron Spin Resonance Spectroscopy) is probably the most important spectroscopic technique for studying chemical structures. Signals of molecules and sub-units of molecules, and the signals of their various spatial arrangements and configurations are much more pronounced in NMR than in any other spectroscopical method. Nevertheless, NMR is still mainly an off-line technique because of elaborate sample preparation requirements, the instruments sensibility to temperature variations, the necessity of instrument tuning (magnet shimming, drift compensation), and the time required to run spectra. However, some recent developments in instrumentation (magnets, radio frequency source and detector circuitry), and the adaptation to flow cells have made NMR spectroscopy applicable to on-line analysis. A major advantage of NMR spectroscopy is the noninvasive character of the technique.

Most of the process measurements so far utilize pulsed NMR in which the free induction decay (FID) signal of magnetic-field aligned molecular spins after the application of a perpendicular radio frequency pulse is analyzed as a function of time. Network formation during the radiation-induced polymerization of *N*-isopropylacrylamide has been monitored by T_2 spin – spin relaxation time measurements [513]. The cure of an epoxy – amine system has been studied by NMR spectroscopy and compared to the results of alternative techniques such as ultrasound and dielectrics [514]. The solids content of rubber latex can be determined without any special sample preparation with a total measurement time of ca. 10 s by quasi-on-line application of pulsed NMR [515]. FID signals in combination with a chemometric technique (PLS) enable one to predict the styrene content in a butadiene – styrene resin [516]. Other applications include determination of hydrogen in various materials, and several solvents in gasoline.

High resolution NMR spectroscopy has been rarely applied to on-line analysis. Because of its natural abundance, ^{1}H NMR is superior to all other nuclei for process measurements. Applications include the determination of the total hydrogen content of hydrocarbons, the determination of crystallinity and melt index of polyethylene, cure dynamics, and copolymer formation monitoring [517], carbamic acid formation in supercritical carbon dioxide [518], the chemical equilibria and reaction kinetics of the formaldehyde – water – methanol system, the reactive gas absorption of CO_2 in aqueous solutions of monoethanolamine [519], are monitoring processes in refineries [520] and the emulsion polymerization of butyl acrylate [521]. Low-field NMR spectroscopy has been shown to be applicable to monitor the production of benzene [522].

[Top of Page]

References

1. C. Hassell, E. M. Bowman, *Appl. Spectrosc.* **52** (1998) 18A. Links
2. D. C. H. Chien, A. Penlidis, *J. Macromol. Sci. – Rev. Macromol. Chem. Phys.* **C30** (1990) 1.
3. A. Penlidis, T. A. Duever, *Macromol. Symp.* **111** (1996) 195. Links
4. K. J. Clevett: *Process Analyzer Technology*, Wiley-Interscience, New York 1986.
5. D. E. Podkulski, *Chem. Eng. Prog.* (1997) no. 10, 33. Links
6. K. G. Carr-Brion, J. R. P. Clarke: *Sampling Systems for Process Analyzers*, Butterworth-Heinemann, London 1996.
7. K. H. Koch: *Industrielle Prozeßanalytik*, Springer-Verlag, Berlin – Heidelberg – New York 1997.
8. G. Oesterle: *Prozeßanalytik — Grundlagen und Praxis*, R. Oldenbourg Verlag, München – Wien 1995.
9. L. Olsson, U. Schulze, J. Nielsen, *Trends Anal. Chem.* **17** (1998) 88.
10. F. McLennan, B. R. Kowalski: *Process Analytical Chemistry*, Blackie Academic & Professional (Chapman & Hall), London 1995.
11. V. Gundelach, L. Litz: *Moderne Prozeßmeßtechnik – Ein Kompendium*, Springer, Berlin 1999.
12. J. M. Chalmers: *Spectroscopy in Process Analysis*, Sheffield Academic Press, Sheffield and CRC Press LLC, Boca Raton 2000.

13. J. E. Puskas, T. E. Long, R. F. Storey: *In Situ Spectroscopy of Monomer and Polymer Synthesis*, Kluwer Academic/Plenum Publ., New York – Dordrecht 2003.
14. K. A. Bakeev: *Process Analytical Technology*, Blackwell Publishing, Oxford 2005.
15. R. W. Kessler: *Prozessanalytik – Strategien und Fallbeispiele aus der industriellen Praxis*, Wiley-VCH, Weinheim, Germany 2006.
16. K. R. Beebe *et al.*, *Anal. Chem.* **65** (1993) 199R. Links
17. W. W. Blaser *et al.*, *Anal. Chem.* **67** (1995) 47R. Links
18. J. Workman *et al.*, *Anal. Chem.* **71** (1999) 121R. Links
19. J. Workman *et al.*, *Anal. Chem.* **73** (2001) 2705. Links
20. J. Workman, M. Koch, D. J. Veltkamp, *Anal. Chem.* **75** (2003) 2859. Links
21. J. Workman, M. Koch, D. J. Veltkamp, *Anal. Chem.* **77** (2005) 3789. Links
22. http://www.infoscience.com/JPAC/.
23. http://www.s-a-s.org/journal/journal.htm.
24. http://pubs.acs.org/journals/ancham/.
25. http://www.springer.com/sgw/cda/frontpage/0,11855,1-135-70-1023710-0,00.html.
26. http://www.ifpac.com/.
27. http://www.cpact.com/apact/apact.html.
28. http://www.iaeac.ch/.
29. http://www.fda.gov/cder/OPS/PAT.htm.
30. http://www.cpac.washington.edu/.
31. http://mcec.engr.utk.edu/.
32. http://www.cpact.com/.
33. http://macc.mcmaster.ca/.
34. http://www.ctac.mis.coventry.ac.uk/.
35. http://www.gdch.de/strukturen/fg/ach/aks/pat__e.htm.
36. H. U. Moritz in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering*, VCH Verlag, Weinheim, Germany 1989, p. 248.
37. R. Riesen in B. F. Mathot (ed.): *Calorimetry and Thermal Analysis of Polymers*, Hanser Publ., München 1994, p. 319.
38. T. Stockhausen, *VDI-Berichte* **No. 429**, VDI-Verlag, Düsseldorf 1996.
39. T. Basalik, *Amer. Lab.* **30** (1998) 33.
40. W. Regenass, *Thermochim. Acta* **95** (1985) 351. Links
41. H. Schuler, C. U. Schmidt, *Chem. Eng. Sci.* **47** (1992) 899. Links
42. R. S. H. Wu, *Chem. Eng. Prog.* **81** (1985) 57.
43. P. Hugo, *Chem. Ing. Tech.* **65** (1993) 1497.
44. L. Guldbaek, J. Villadsen, *Chem. Eng. Sci.* **42** (1987) 1153.
45. F. Groenlund, *Thermochim. Acta* **175** (1991) 63. Links
46. D. Dollimore, S. Lerdkanchanaporn, *Anal. Chem.* **70** (1998) 27R. Links
47. L. Verala de la Rosa, E. D. Sudol, M. S. El-Aasser, A. Klein, *J. Polym. Sci., A: Polym. Chem.* **34** (1996) 461. Links
48. R. Carloff, A. Proß, K.-H. Reichert, *Chem. Eng. Tech.* **17** (1994) 406.
49. A. Tietze, A. Proß, K.-H. Reichert, *DECHEMA Monogr.* **131** (1995) 673.
50. P.G. De Luca, C. Scali, *Chem. Eng. Sci.* **57** (2002) 2077. Links
51. R. Gesthuisen, S. Krämer, G. Niggemann, J. R. Leiza, J. M. Asua, *Comp. Chem. Eng.* **29** (2005) 349.
52. Philips Petroleum, US 3 275 809, 1963 (E. D. Tolin, D. A. Fluegel).
53. ICI, UK 1 217 325, 1967 (B. N. Hendy).
54. ICI, UK 1 549 841, 1975 (B. F. Beckingham, B. N. Hendy, J. V. Simons).
55. DOW, DE 2 834 569, 1977 (L. C. Walker).
56. BASF, US 6 498 219, 2002 (H. Kröner, R. Klostermann, J. Birk, T. Hauff).
57. BASF, DE 102 25 383, 2002 (U. Nieken, J.R. von Watzdorf, J. Kessler, C. Benthack, H. Kröner, T. Hauff, P. Schlemmer).
58. M. Mosebach, K.-H. Reichert, *J. Appl. Polym. Sci.* **66** (1997) 673. Links
59. A. Pross, K. Platkowski, K.-H. Reichert, *Polym. Intl.* **45** (1998) 22.
60. A. Urretabizkaia, E. D. Sudol, M. S. El-Aasser, J. M. Asua, *J. Polym. Sci., A: Polym. Chem.* **31** (1993) 2907. Links
61. J. R. Infante *et al.*, *J. Appl. Polym. Sci.* **62** (1996) 2311. Links
62. A. Urretabizkaia, J. R. Leiza, J. M. Asua, *AIChE J.* **40** (1994) 1850. Links
63. I. S. de Buruaga *et al.*, *AIChE J.* **43** (1997) 1069. Links
64. O. Ubrich, B. Srinivasan, P. Lerena, D. Bonvin, F. Stoessel, *Chem. Eng. Sci.* **56** (2001) 5147. Links
65. G. Hessel *et al.*, *Chem Ing. Tech.* **74** (2002) 1692.
66. S. BenAmor, D. Colombie, T. McKenna, *Ind. Eng. Chem. Res.* **41** (2002) 4233. Links
67. S. Fan, E. Alpay, *Chem. Eng. Sci.* **59** (2004) 2811.

68. F. Stoessel, *Chem. Eng. Progr.* **89** (1993) 68.
69. A. Grefenstein: *Reaktive Extrusion und Aufbereitung*, Carl Hanser Verlag, München, Wien 1996.
70. C. W. Macosko: *RIM — Fundamentals of Reaction Injection Molding*, Hanser Publ., München – Wien – New York 1989.
71. BASF, WO 03/044 607, 2002 (M. Taylor, J. Antonucci, R. Racz).
72. D. S. Giordani, L. M. F. Lona, T. F. McKenna, M. A. Krähenbühl, A. M. dos Santos, *Macromol. Mat. Eng.* **290** (2005) 485.
73. A. F. Santos, J. C. Pinto, T. F. McKenna, *Macromol. Symp.* **226** (2005) 157.
74. D. J. am Ende, P. J. Clifford, D. M. DeAntonis, C. SantaMaria, S. J. Brenek, *Org. Proc. Res. Dev.* **3** (1999) 319.
75. A. Zogg, U. Fischer, K. Hungerbühler, *Chem. Eng. Sci.* **59** (2004) 5795; Links A. Zogg, U. Fischer, K. Hungerbühler, *Chemometr. Intell. Lab. Syst.* **71** (2004) 165.
76. O. Elizalde, M. Azpeitia, M. M. Reis, J. M. Asua, J. R. Leiza, *Ind. Eng. Chem. Res.* **44** (2005) 7200. Links
77. K. S. Anseth, R. A. Scott, N. A. Peppas, *Macromolecules* **29** (1996) 8308. Links
78. B. Sandner, N. Kotzian, J. Tübke, S. Wartewig, O. Lange, *Macromol. Chem. Phys.* **198** (1997) 2715. Links
79. R. A. Scott, N. A. Peppas, *AIChE J.* **43** (1997) 135. Links
80. J. L. Han, Y. C. Chern, K. H. Hsieh, W. Y. Chiu, C. C. M. Ma, *J. Appl. Polym. Sci.* **68** (1998) 121. Links
81. P. Hauptmann, R. Lucklum, B. Henning in H. Baltes, W. Göpel, J. Hesse (eds.): *Sensors Update*, **vol. 3**, Wiley-VCH Verlag GmbH, Weinheim, Germany 1998, p. 163.
82. L. C. Lynnworth: *Ultrasonic Measurement for Process Control*, Academic Press, New York 1989.
83. J. W. Grate, G. C. Frye in H. Baltes, W. Göpel, J. Hesse (eds.): *Sensors Update*, **vol. 2**, Wiley-VCH Verlag GmbH, Weinheim, Germany 1997, p. 37.
84. K. S. Suslick (ed.): *Ultrasound: Its Chemical Physical and Biological Effects*, VCH Publishers, New York 1988.
85. P. Hauptmann: *Sensoren — Prinzipien und Anwendungen*, Hanser Verlag, München 1991.
86. B. Hartmann in R. A. Fava (ed.): *Methods of Experimental Physics*, **vol. 16c**, Academic Press, New York 1980, p. 59.
87. J. P. Lorimer, T. J. Mason: *Sonochemistry: Theory, Applications and Uses of Ultrasound in Chemistry*, Ellis Horwood, Chichester 1988.
88. T. J. Mason: *Sonochemistry: The Uses of Ultrasound in Chemistry*, The Royal Society of Chemistry, Cambridge 1990.
89. P. Sladky, I. Pelant, L. Parma, *Ultrasonics* **16** (1979) 32.
90. P. Sladky, L. Parma, I. Zdrazil, *Polym. Bull.* **7** (1982) 401. Links
91. H. G. Lüscher, *Kunststoffe-Plastics* **23** (1976) 10, 16.
92. E. P. Papadakis, *J. Appl. Phys.* **45** (1974) 1218. Links
93. F. Dinger, P. Hauptmann, R. Säuberlich, S. Wartewig, *Acta Polym.*, **34** (1983) 303. Links
94. P. Hauptmann, F. Dinger, R. Säuberlich, *Polymer*, **26** (1985) 1741. Links
95. W.-D. Hergeth, P. Starre, K. Schmutzler, S. Wartewig, *Polymer*, **29** (1988) 1323. Links
96. S. Canegallo, M. Apostolo, G. Storti, M. Morbidelli, *J. Appl. Polym. Sci.* **57** (1995) 1333. Links
97. A. Siani, M. Apostolo, M. Morbidelli, *DECHEMA Monogr.* **131** (1995) 149.
98. G. Storti, A. K. Hipp, M. Morbidelli, *Polym. React. Eng.* **8** (2000) 77. Links
99. G. S. Constable, A. J. Lesser, E. B. Coughlin, *J. Polym. Sci, B: Polym. Phys.* **41** (2003) 1323. Links
100. I. Alig, M. Fedtke, K.-G. Häusler, W. Tänzer, S. Wartewig, *Progr. Colloid Polym. Sci.* **78** (1988) 54.
101. I. Alig, K.-G. Häusler, W. Tänzer, S. Unger, *Acta Polym.* **39** (1988) 269. Links
102. I. Alig, K.-G. Häusler, K. Nancke, E. Domaratius, M. Fedtke, *Acta Polym.* **40** (1989) 590. Links
103. I. Alig, D. Lellinger, K. Nancke, A. Rizos, G. Fytas, *J. Appl. Polym. Sci.* **44** (1992) 829. Links
104. M. Younes, S. Wartewig, D. Lellinger, B. Strehmel, V. Strehmel, *Polymer* **35** (1994) 5269. Links
105. I. Alig, K. Nancke, G. P. Johari, *J. Polym. Sci., B: Polym. Phys.* **32** (1994) 1465. Links
106. J.-F. Zhou, G. P. Johari, *Macromolecules* **30** (1997) 8085. Links
107. M. Petrick, *Thesis,* RWTH Aachen, 1975.
108. R. Gendron, J. Tatibouët, J. Guévremont, M. M. Dumoulin, L. Piché, *Polym. Eng. Sci.* **35** (1995) 79. Links
109. I. Alig, D. Lellinger, R. Lamour, J. Ramthun, *Kunststoffe-Plastics* (in press).
110. P. D. Coates, S. E. Barnes, M. G. Sibley, E. C. Brown, H. G. M. Edwards, I. J. Scowen, *Polymer* **44** (2003) 5937. Links
111. S. E. Barnes, E. C. Brown, M. G. Sibley, H. G. M. Edwards, I. J. Scowen, P. D. Coates, *Appl. Spectrosc.* **59** (2005) 611. Links
112. A. M. Howe, A. R. Mackie, M. M. Robins, *J. Disp. Sci. Technol.* **7** (1986) 231.
113. D. Hibberd *et al.*, *J. Colloid Interface Sci.* **193** (1997) 77. Links
114. S. Biggs, F. Grieser, *Macromolecules* **28** (1995) 4877. Links
115. G. Cooper, F. Grieser, S. Biggs, *J. Colloid Interface Sci.* **184** (1996) 52. Links
116. M. Romdhane, A. Gadri, F. Contamine, C. Gourdon, G. Casamatta, *Ultrasonics Sonochem.* **4** (1997) 235.
117. D. J. McClements, *Adv. Colloid Interface Sci.* **37** (1991) 33. Links
118. D. Roberts, *International Laboratory News*, June 1996.

119. A. S. Dukhin, P. J. Goetz, *Langmuir* **21** (1996) 4336.
120. A. S. Dukhin, P. J. Goetz, *Langmuir* **12** (1996) 4987. Links
121. A. S. Dukhin, P. J. Goetz, C. W. Hamlet, *Langmuir* **12** (1996) 4998. Links
122. R. W. O'Brien, D. W. Cannon, W. N. Rowlands, *J. Colloid Interface Sci.* **173** (1995) 406.
123. M. Stintz, F. Hinze, S. Ripperger, *Environm. Protect. Eng.* **25** (1999) 15.
124. F. Hinze, S. Ripperger, M. Stintz, *Chem. Ing. Tech.* **72** (2000) 322.
125. F. Babick, S. Ripperger, *Part. Part. Syst. Charact.* **19** (2002) 176. Links
126. L. P. Adjadj, G. Storti, M. Morbidelli, *Langmuir* **19** (2003) 3935.
127. F. Babick, S. Ripperger, *Chem. Ing. Tech.* **76** (2004) 30.
128. Z. Lin, M. D. Ward, *Anal. Chem.* **68** (1996) 1285. Links
129. A. Weerawardena, C. J. Drummond, F. Caruso, M. McCormick, *Langmuir* **14** (1998) 575. Links
130. S. J. Martin, A. J. Ricco, *Sensors Actuators* **A21 – A23** (1990) 712.
131. A. J. Ricco, A. W. Staton, R. M. Crooks, T. Kim, *Faraday Discuss.* **107** (1997) 247.
132. J. Dorighi, S. Krishnaswamy, J. D. Achenbach, *Rev. Progr. Quant. Nondestruct. Eval.* **15** (1996) 657.
133. B. Mitra, D. J. Booth, *Ultrasonics* **35** (1998) 569. Links
134. J. W. R. Boyd, J. Varley, *AIChE J.* **44** (1998) 1731; Links J. W. R. Boyd, J. Varley, *Chem. Eng. Sci.* **56** (2001) 1749. Links
135. S. Havriliak, S. J. Havriliak: *Dielectric and Mechanical Relaxation in Materials*, Hanser Publishers, Munich – Vienna – New York 1997.
136. J. R. Macdonald: *Impedance Spectroscopy*, John Wiley & Sons, New York 1987.
137. G. Williams, *Adv. Polym. Sci.* **33** (1979) 60.
138. A. K. Jonscher: *Dielectric Relaxation in Solids*, Chelsea Dielectrics Press, London 1983.
139. J. Mijovic, F. Bellucci, *TRiP* **4** (1996) 74.
140. K. Funke, C. Cramer, *Curr. Opin. Solid State Mat. Sci.* **2** (1997) 483.
141. A. Rost: *Messung dielektrischer Stoffeigenschaften*, Akademie-Verlag, Berlin 1978.
142. F. Kremer, M. Arndt in J. P. Runt, J. J. Fitzgerald (eds.): *Dielectric Spectroscopy of Polymeric Materials*, American Chemical Society, Washington 1997, p. 67.
143. D. A. Wasylyshyn, M. G. Parthun, G. P. Johari, *J. Mol. Liq.* **69** (1996) 283. Links
144. D. E. Kranbuehl in J. P. Runt, J. J. Fitzgerald (eds.): *Dielectric Spectroscopy of Polymeric Materials*, American Chemical Society, Washington D.C. 1997, p. 303.
145. G. Williams, C. Duch, J. Fourier, J. R. Hayden in A. H. Fawkett (ed.): *Polymer Spectroscopy*, John Wiley & Sons, Chichester 1996, p. 275.
146. S. Andjelic, J. Mijovic, F. Bellucci, *J. Polym. Sci., B: Polym. Phys.* **36** (1998) 641. Links
147. S. Andjelic, J. Mijovic, *Macromolecules* **31** (1998) 2872. Links
148. E. Tombardi, C. Ferrari, G. Salvetti, G. P. Johari, *J. Polym. Sci., B: Polym. Phys.* **36** (1998) 303. Links
149. I. Alig, W. Jenninger, M. Junker, L. A. de Graaf, *J. Macromol. Sci.-Phys.* **B35** (1996) 563.
150. M. Fedtke, F. Schramm, *KGK Kautschuk Gummi Kunststoffe* **51** (1998) 201.
151. P. Bartolomeo, J. F. Chailan, J. L. Vernet, *Polymer* **42** (2001) 4385. Links
152. D. Schuele, R. Renner, D. Coleman, *Mol. Cryst. Liq. Cryst.* **299** (1997) 343. Links
153. A. S. Skordos, I. K. Partidge, *J. Polym. Sci. B: Polym. Phys.* **42** (2004) 146. Links
154. J. W. Kwon, W. S. Chin, D. G. Lee *J. Adhesion Sci. Technol.* **17** (2003) 2111.
155. E. H. B. deLacey, L. R. White, *J. Chem. Soc. Faraday Trans. 2* **77** (1981) 2007.
156. R. M. Fitch, L. S. Su, S. L. Tsaur in M. S. El-Aasser, R. M. Fitch (eds.): *Future Directions in Polymer Colloids*, Martinus Nijhoff, Dordrecht – Boston – Lancaster, 1987.
157. R. M. Fitch, L. S. Su, S. L. Tsaur in F. Candau, R. H. Ottewil (eds.): *Scientific Methods for the Study of Polymer Colloids and Their Applications*, Kluwer Academic, Dordrecht 1990.
158. L. S. Su, S. Jayasuriya, R. M. Fitch in J. W. Goodwin, R. Buscall (eds.): *Colloid Polymer Particles*, Academic Press, London 1995.
159. F. Carrique, A. V. Delgado, *Colloid Polym. Sci.* **274** (1996) 848. Links
160. A. A. Garrouch, H. M. S. Lababidi, R. B. Gharbi, *J. Phys. Chem.* **100** (1996) 16996.
161. Y. Chonde, M. Shabrang, *J. Colloid Interface Sci.* **186** (1997) 248. Links
162. B. B. Sauer, R. S. Stock, K.-H. Lim, W. H. Ray, *J. Appl. Polym. Sci.* **39** (1990) 2419. Links
163. J. M. Harnly, R. E. Fields, *Appl. Spectrosc.* **51** (1997) 334A. Links
164. C. W. Brown in F. M. Mirabella: *Modern Techniques in Applied Molecular Spectroscopy*, John Wiley & Sons, New York 1998, p. 377.
165. B. Schrader: *Infrared and Raman Spectroscopy*, VCH, Weinheim, Germany 1995.
166. K. Vedam, *Thin Solid Films* **313–314** (1998) 1.
167. R. W. Collins, I. An, H. Fujiwara, J. Lee, Y. Lu, J. Koh, P. I. Rovira, *Thin Solid Films* **313–314** (1998) 18.
168. S. P. Gurden, J. A. Westerhuis, A. K. Smilde, *AIChE J.* **48** (2002) 2283.
169. G. Rinke, U. Hoeppener-Kramar, *GIT Fachz. Lab.* **10/95**, 939.
170. A. C. Quinn, P. J. Gemperline, B. Baker, M. Zhu, D. S. Walker, *Chemometr. Intell. Lab. Syst.* **45** (1999) 199.

171. F. Hunt, H. Ohde, C.M. Wai, *Rev. Sci. Instr.* **70** (1999) 4661.
172. G. Vishnoi, T. C. Goel, P. K. C. Pillai, *Sensors Actuators* **B 45** (1997) 43.
173. S. A. Merschman, D. C. Tilotta, *Appl. Spectrosc.* **52** (1998) 106. Links
174. J. C. Phelan, C. S. P. Sung, *Macromolecules* **30** (1997) 6845. Links
175. H. J. Paik, N. H. Sung, C. S. P. Sung, *Polym. Prep.* **32** (1991) 669.
176. J. B. Reeves, *J. Near Infrared Spectrosc.* **2** (1994) 199.
177. P. J. Gemperline, *Chemotr. Intell. Lab. Syst.* **39** (1997) 29.
178. E. Bouveresse, C. Hartmann, D. L. Massart, I. R. Last, K. A. Prebble, *Anal. Chem.* **68** (1996) 982. Links
179. J. Workmann, *J. Near Infrared Spectrosc.* **1** (1993) 221.
180. W. F. McClure, *Anal. Chem.* **66** (1994) 43A. Links
181. D. Noble, *Anal. Chem.* **67** (1995) 735A. Links
182. M. Xanthos, S. H. Patel, *Adv. Polym. Tech.* **14** (1995) 151. Links
183. D. E. Podkulski, *Chem. Eng. Progr.* (1997) Oct. 33.
184. N. Poisson, G. Lachenal, H. Sautereau, *Vib. Spectrosc.* **12** (1996) 237. Links
185. C. Billaud, M. Vandeuren, R. Legras, V. Carlier, *Appl. Spectrosc.* **56** (2002) 1413. Links
186. J. Mijovic, S. Andjelic, *Polymer* **36** (1995) 3783. Links
187. J. Mijovic, S. Andjelic, J. M. Kenny, *Polym. Adv. Technol.* **7** (1996) 1. Links
188. J. P. Dunkers, K. M. Flynn, M. T. Huang, W. G. McDonough, *Appl. Spectrosc.* **52** (1998) 552. Links
189. L. Xu, J. R. Schlup, *Appl. Spectrosc.* **50** (1996) 109. Links
190. H.-R. Bjørsvik, *Appl. Spectrosc.* **50** (1996) 1541.
191. E. D. Yalvac, M. B. Seasholtz, M. A. Beach, S. R. Crouch, *Appl. Spectrosc.* **51** (1997) 1565. Links
192. C. Wu, J. D. S. Danielson, J. B. Callis, M. T. Eaton, N. L. Ricker, *Process Control Qual.* **8** (1996) 25.
193. S. Schönkopf, *Chem. Tech.* **27** (1998) 10.
194. C. Coffey, A. Predoehl, D. S. Walker, *Appl. Spectrosc.* **52** (1998) 717. Links
195. T. Ogasawara, A. Nara, H. Okabayashi, E. Nishio, C. J. O'Connor, *Colloid Polym. Sci.* **278** (2000) 946.
196. A. Olinga, R. Winzen, H. Rehage, H. W. Siesler, *J. Near Infrared Spectrosc.* **9** (2001) 19.
197. R. A. M. Vieira, C. Sayer, E. L. Lima, J. C. Pinto, *J. Appl. Polym. Sci.* **84** (2002) 2670. Links
198. H. H. A. Lousberg, H. F. M. Boelens, E. P. le Comte, H. C. J. Hoefsloot, A. K. Smilde, *J. Appl. Polym. Sci.* **84** (2002) 90. Links
199. A. Cherfi, G. Fevotte, *Macromol. Chem. Phys.* **203** (2002) 1188. Links
200. J. M. R. Fontoura, A. F. Santos, F. M. Silva, M. K. Lenzi, E. L. Lima, J. C. Pinto, *J. Appl. Polym. Sci.* **90** (2003) 1273. Links
201. F. H. Florenzano, P. Enohnyaket, V. Fleming, W. F. Reed, *Eur. Polym. J.* **41** (2005) 535. Links
202. N. Sheibat-Othman, D. Peycelon, J.-B. Egraz, J.-M. Suau, G. Fevotte, *NIR news* **16** (2005) 8.
203. R. A. M. Vieira, C. Sayer, E. L. Lima, J. C. Pinto, *Ind. Eng. Chem. Res.* **41** (2002) 2915. Links
204. A. Cherfi, G. Fevotte, C. Novat, *J. Appl. Polym. Sci.* **85** (2002) 2510. Links
205. N. Sheibat-Othman, G. Fevotte, D. Peycelon, J.-B. Egraz, J.-M. Suau, *AIChE J.* **50** (2004) 654. Links
206. H. W. Ward II *et al.*, *Appl. Spectrosc.* **52** (1998) 17. Links
207. J. Rantanen, H. Wikström, R. Turner, L. S. Taylor, *Anal. Chem.* **77** (2005) 556. Links
208. R. L. Green, G. Thurau, N. C. Pixley, A. Mateos, R. A. Reed, J. P. Higgins, *Anal. Chem.* **77** (2005) 4515. Links
209. F. Cuesta Sanchez *et al.*, *Fresenius J. Anal. Chem.* **352** (1995) 771. Links
210. S. S. Sekulic *et al.*, *Anal. Chem.* **68** (1996) 509. Links
211. R. de Maesschalck, F. Cuesta Sanchez, D. L. Massart, P. Doherty, P. Hailey, *Appl. Spectrosc.* **52** (1998) 725. Links
212. R. M. Mosley, R. R. Williams, *J. Near Infrared Spectrosc.* **2** (1994) 119.
213. H. F. M. Boelens, W. T. Kok, O. E. de Noord, A. K. Smilde, *Appl. Spectrosc.* **54** (2000) 406. Links
214. H. Chung, M.-S. Ku, *Vib. Spectrosc.* **31** (2003) 125. Links
215. J. P. Higgins *et al.*, *Anal. Chem.* **75** (2003) 1777. Links
216. A. Szalay *et al.*, *Part. Part. Syst. Charact.* **22** (2005) 219. Links
217. M. M. Reis, P. H. H. Araujo, *Macromol. Rapid Commun.* **24** (2003) 620. Links
218. W. H. A. M. van den Broek *et al.*, *Chemometr. Intell. Lab. Syst.* **35** (1996) 187.
219. W. H. A. M. van den Broek, D. Wienke, W. J. Melssen, L. M. C. Buydens, *Anal. Chim Acta* **361** (1998) 161. Links
220. W. H. A. M. van den Broek *et al.*, *Appl. Spectrosc.* **51** (1997) 856. Links
221. M. Watari, Y. Ozaki, *Appl. Spectrosc.* **59** (2005) 912. Links
222. D. Fischer, T. Bayer, K.-J. Eichhorn, M. Otto, *Fresenius J. Anal. Chem.* **359** (1997) 74. Links
223. M. Stephan, O. Franzheim, D. Fischer in *Jahresbericht 1997*, IPF Institut für Polymerforschung Dresden e.V., Dresden 1997.
224. L. Jakisch, D. Fischer, M. Stephan, K.-J. Eichhorn, *Kunststoffe* **85** (1995) 1338.
225. M. G. Hansen, A. Khettry, *Tappi J.* **78** (1995) 129. Links
226. A. Khettry, M. G. Hansen, *Polym. Eng. Sci.* **36** (1996) 1232. Links
227. M. G. Hansen, S. Vedula, *J. Appl. Polym. Sci.* **68** (1998) 859. Links

228. S. Vedula, M. G. Hansen, *J. Appl. Polym. Sci.* **68** (1998) 873. Links
229. A. Hirsch, B. E. Bridgland, *Can. Spectrosc.* May 1967, 101.
230. E. G. Chatzi, O. Kammona, C. Kiparissides, *J. Appl. Polym. Sci.* **63** (1997) 799. Links
231. C. Decker, K. Moussa, *Macromolecules* **22** (1989) 4455. Links
232. J. E. Dietz, B. J. Elliot, N. A. Peppas, *Macromolecules* **28** (1995) 5163. Links
233. H. Hua, M. A. Dube, *Polymer* **42** (2001) 6009. Links
234. H. Hua, M. A. Dube, *J. Polym. Sci., A: Polym. Chem.* **39** (2001) 1860. Links
235. J. E. Puskas, M. G. Lanzendörfer, W. E. Pattern, *Polym. Bull.* **40** (1998) 55. Links
236. R. F. Storey, A. B. Donnalley, T. L. Maggio, *Macromolecules* **31** (1998) 1523. Links
237. P. MacLaurin, N. C. Crabb, I. Wells, P. J. Worsfold, D. Coombs, *Anal. Chem.* **68** (1996) 1116. Links
238. A. H. Ali, K. S. V. Srinivasan, *J. Appl. Polym. Sci.* **67** (1998) 441. Links
239. X. Fang, D. F. Rogers, D. A. Scola, M. P. Stevens, *J. Polym. Sci., A: Polym. Chem.* **36** (1998) 461. Links
240. A. M. Kaminski, M. W. Urban, *J. Coat. Technol.* **69** (1997) 113.
241. S. J. Martin, V. J. McBrierty, D. C. Douglas, *Macromolecules* **34** (2001) 8934. Links
242. T. Scherzer, A. Tauber, R. Mehnert, *Vib. Spectrosc.* **29** (2002) 125. Links
243. S. J. Oh, J. L. Koenig, *Anal. Chem.* **70** (1998) 1768. Links
244. T. Kimura, H. Ezure, S. Tanaka, E. Ito, *J. Polym. Sci., B: Polym. Phys.* **36** (1998) 1227. Links
245. D. Qin, G. Cadet, *Anal. Chem.* **69** (1997) 1942. Links
246. A. Beenen, R. Niessner, *Analyst* **123** (1998) 543. Links
247. M. W. Urban, E. M. Salazar-Rojas, *J. Polym. Sci., A: Polym. Chem.* **28** (1990) 1593. Links
248. T. Heitz, B. Drevillon, J. E. Bouree, C. Godet, *Appl. Phys. Letts.* **72** (1998) 780.
249. G. Mazarevica, J. Diewok, J. R. Baena, E. Rosenberg, B. Lendl, *Appl. Spectrosc.* **58** (2004) 804. Links
250. A. Pintar, J. Batista, J. Levec, *Analyst* **127** (2002) 1535. Links
251. P. B. Graham, J. D. Bergevin, *Process Control Qual.* **5** (1993) 145.
252. K. de Braekeleer *et al.*, *Appl. Spectrosc.* **54** (2000) 601.
253. L. M. Bodecchi, M. Cocchi, M. Malagoli, M. Manfredini, A. Marchetti, *Anal. Chim. Acta* **554** (2005) 207. Links
254. P. Marteau, F. Adar, N. Zanier-Szydlowski, *Amer. Lab.* **28** (1996) 21H.
255. J. M. Chalmers, N. J. Everall *Trends Anal. Chem.* **15** (1996) 18.
256. B. Chase, *Appl. Spectrosc.* **48** (1994) 14A. Links
257. B. Schrader, A. Hoffmann, A. Simon, J. Sawatzki, *Vib. Spectrosc.* **1** (1991) 239. Links
258. S. M. Angel, M. Carrabba, T. F. Cooney, *Spectrochim. Acta* **A51** (1995) 1779.
259. I. R. Lewis, P. R. Griffiths, *Appl. Spectrosc.* **50** (1996) 12A. Links
260. C. K. Chong *et al.*, *Vib. Spectrosc.* **3** (1992) 35. Links
261. T. F. Cooney, H. T. Skinner, S. M. Angel, *Appl. Spectrosc.* **50** (1996) 836, 849. Links
262. B. Chase, *Microchim. Acta* **14** (1997) 1.
263. F. Adar, R. Geiger, J. Noonan, *Appl. Spectrosc. Rev.* **32** (1997) 45. Links
264. D. Ben-Amotz, F. LaPlant, Y. Jiang, T. Biermann, *Proceedings 35th Ann. Polyurethane Technical/Marketing Conference 1994*, 375.
265. C. Henry, *Anal. Chem. News Feat.* 1997, 309A.
266. M. Pelletier, K. Davis, *Amer. Lab.* **28** (1996) 34C.
267. M. B. Denton, D. A. Gilmore, *Proc. SPIE – Int. Soc. Opt. Eng.* **2388** (1995) 121.
268. K. Witke, W. Kimmer, *Plaste & Kautschuk* **23** (1976) 799.
269. K. Witke, H.-G. Buge, K.-W. Brzezinka, W. Kimmer, *Acta Polym.* **34** (1983) 627. Links
270. B. Chu, G. Fytas, G. Zalczer, *Macromolecules* **14** (1981) 395. Links
271. W. M. Sears, J. L. Hunt, J. R. Stevens, *J. Chem. Phys.* **75** (1981) 1589, 1599. Links
272. E. Gulari, K. McKeigue, K. Y. S. Ng, *Macromolecules* **17** (1984) 1822. Links
273. S. Damoun, R. Papin, G. Ripault, M. Rousseau, J. C. Rabadeux, D. Durand, *J. Raman Spectrosc.* **23** (1992) 385. Links
274. M. Kamath *et al.*, *Macromolecules* **26** (1993) 5954. Links
275. L. Feng, K. Y. S. Ng, *Macromolecules* **23** (1990) 1048. Links
276. C. Wang, T. J. Vickers, J. B. Schlenoff, C. K. Mann, *Appl. Spectrosc.* **46** (1992) 1729, and Links *Appl. Spectrosc.* **47** (1993) 928. Links
277. T. J. Vickers, C. K. Mann, *Proc. SPIE – Int. Soc. Opt. Eng.* **2367** (1995) 219, and *Proc. SPIE – Int. Soc. Opt. Eng.* **2504** (1995) 310.
278. T. Özpozan, B. Schrader, S. Keller, *Spectrochim. Acta* **A53** (1997) 1.
279. A. Brookes, J. M. Dyke, P. J. Hendra, A. Strawn, *Spectrochim. Acta* **A53** (1997) 2302.
280. J. Haigh, A. Brookes, P. J. Hendra, A. Strawn, C. Nicholas, M. Purbrick, *Spectrochim. Acta* **A53** (1997) 9.
281. A. N. Shchegolikhin, O. L. Lazareva, *Spectrochim. Acta* **A53** (1997) 67.
282. O. Elizalde, J. M. Asua, J. R. Leiza, *Appl. Spectrosc.* **59** (2005) 1270, 1280. Links
283. M. M. Reis, P. H. H. Araujo, C. Sayer, R. Giudici, *Macromol. Symp.* **206** (2004) 165. Links

284. C. Q. Yang, X. Gu, *J. Appl. Polym. Sci.* **81** (2001) 223. Links
285. M. van den Brink, M. Pepers, A. M. van Herk, A. L. German, *Polym. React. Eng.* **9** (2001) 101.
286. J. B. Cooper, T. M. Vess, L. A. Campbell, B. J. Jensen, *J. Appl. Polym. Sci.* **62** (1996) 135. Links
287. S. M. Angel, M. L. Myrick, *Proc. Ann. Meet. Adhes. Soc.* **19** (1996) 379.
288. J. F. Aust, K. S. Booksh, C. M. Stellmann, R. S. Parnas, M. L. Myrick, *Appl. Spectrosc.* **51** (1997) 247. Links
289. M. Ghosh, S. Chakrabarti, T. N. Misra, *J. Raman Spectrosc.* **29** (1998) 263. Links
290. W. Kessler, G. Luft, W. Zeiß, *Ber. Bunsenges. Phys. Chem.* **101** (1997) 698.
291. W. C. Bell, K. S. Booksh, M. L. Myrick, *Anal. Chem.* **70** (1998) 332. Links
292. P. Jacob, B. Wehling, W. Hill, D. Klockow, *Appl. Spectrosc.* **51** (1997) 74. Links
293. M. Cakmak, F. T. Serhatkulu, M. Graves, J. Galay, *ANTEC '97*, 1794.
294. G. J. Gervasio, M. J. Pelletier, *AT-Proc.* **1 – 2** (1997) 7.
295. M. L. Myrick, S. M. Angel, M. Gallagher, *21st Conf. Fed. Anal. Chem. Spectrosc. Soc.*, St. Louis 1994, no. 220.
296. O. Svensson, M. Josefson, F. W. Langkilde, *Chemometr. Intell. Lab. Syst.* **49** (1999) 49.
297. I. A. Clegg, N. J. Everall, B. King, H. Melvin, C. Norton, *Appl. Spectrosc.* **55** (2001) 1138. Links
298. A. D. Shaw *et al.*, *Appl. Spectrosc.* **53** (1999) 1419. Links
299. P. J. Aarnoutse, J. A. Westerhuis, *Anal. Chem.* **77** (2005) 1228. Links
300. E. D. Lipp, R. L. Grosse, *Appl. Spectrosc.* **52** (1998) 42. Links
301. J. F. Lübben, B. Schrader, *J. Mol. Struct.* **410 – 411** (1997) 543.
302. C. L. Aardahl, J. F. Widmann, E. J. Davis, *Appl. Spectrosc.* **52** (1998) 47. Links
303. E. J. Davis, C. L. Aardahl, J. F. Widmann, *J. Dispersion Sci. Technol.* **19** (1998) 293.
304. J. Musick, J. Popp, M. Trunk, W. Kiefer, *Appl. Spectrosc.* **52** (1998) 692. Links
305. K. Ito, T. Kato, T. Ona, *J. Raman Spectrosc.* **33** (2002) 466. Links
306. M. M. Reis, P. H. H. Araujo, C. Sayer, R. Giudici, *Polymer* **44** (2003) 6123. Links
307. M. Schäfer, R. Starzmann, A. H. Foitzik, *Int. J. Comp. Eng. Sci.* **4** (2003) 601.
308. L. G. Weyer, D. Yee, G. I. Pearson, J. Wade, *Appl. Spectrosc.* **58** (2004) 1500. Links
309. T. R. McCaffery, Y. G. Durant, *J. Appl. Polym. Sci.* **86** (2002) 1507. Links
310. M. Kuball, *Surface Interface Anal.* **31** (2001) 987.
311. L. K. Ballast, R. W. Bormett in D. G. Seiler *et al.* (eds.): *Characterization and Metrology for ULSI Technology*, American Institute of Physics, Melville N.Y. 2001, pp. 327.
312. P. He, S. Bitla, D. Bousfield, C. P. Tripp, *Appl. Spectrosc.* **56** (2002) 1115. Links
313. H. Wikstrom, I. R. Lewis, L. S. Taylor, *Appl. Spectrosc.* **59** (2005) 934.
314. M. W. Urban: *Vibrational Spectroscopy of Molecules and Macromolecules on Surfaces*, John Wiley & Sons, New York 1993.
315. A. Ruperez, J. J. Laserna in J. J. Laserna (ed.): *Modern Techniques in Raman Spectroscopy*, John Wiley & Sons, New York 1996, p. 227.
316. D. L. Stokes, J. P. Alarie, T. Vo-Dinh, *Proc. SPIE – Int. Soc. Opt. Eng.* **2504** (1995) 552.
317. W. Hill, B. Wehling, C. G. Gibbs, C. D. Gutsche, D. Klockow, *Anal. Chem.* **67** (1995) 3187. Links
318. W. Hill, B. Wehling, V. Fallourd, D. Klockow, *Spectrosc. Europe* **7** (1995) 20.
319. R. Kellner, B. Mizaikoff, M. Jakusch, H. D. Wanzenböck, N. Weissenbacher, *Appl. Spectrosc.* **51** (1997) 495. Links
320. E. P. C. Lai, H. S. Ghaziaskar, *Appl. Spectrosc.* **48** (1994) 1011. Links
321. J. R. Lakowicz: *Principles of Fluorescence Spectroscopy*, Kluwer Academic, Dordrecht 1999.
322. A. Sharma, S. G. Schulmann: *Introduction to Fluorescence Spectroscopy*, John Wiley & Sons, New York 1999.
323. S. R. Shield, J. M. Harris, *Anal. Chem.* **70** (1998) 2576. Links
324. P. Douša, C. Koňak, V. Fidler, K. Dušek, *Polym. Bull.* **22** (1989) 585.
325. J. C. Song, D. C. Neckers, *Polym. Eng. Sci.* **36** (1996) 394.
326. J. L. Jessop, A. B. Scranton, G. J. Blanchard, *Polym. Mat. Sci. Eng.* **72** (1995) 58.
327. F. Mikes, J. Gonzales-Benito, J. Baselga Llido, *J. Polym. Sci B: Polym. Phys.* **42** (2004) 64.
328. L. T. Pearson, P. E. J. Louis, R. G. Gilbert, D. H. Napper, *J. Polym. Sci., A: Polym. Chem.* **29** (1991) 515. Links
329. M. A. Winnik, H. Xu, R. Satguru, *Macromol. Chem., Macromol. Symp.* **70/71** (1993) 107.
330. I. Lacik, J. Bartoň, G. G. Warr, *Macromol. Chem. Phys.* **196** (1995) 2223. Links
331. Z. J. Wang, J. C. Song, R. Bao, D. C. Neckers, *J. Polym. Sci. B Polym. Phys.* **34** (1996) 325. Links
332. M.-A. Chung, J. Adams, J. Fuhrmann, *Polym. Bull.* **40** (1998) 195. Links
333. S. Rudschuk, J. Adams, J. Fuhrmann, *Macromol. Chem. Phys.* **199** (1998) 771. Links
334. D. R. Kim, D. H. Napper, *Macromol. Rapid Commun.* **17** (1996) 845. Links
335. M. A. Winnik, M. D. Croucher in M. S. El-Aasser, R. M. Fitch (eds.): *Future Directions in Polymer Colloids*, Martinus Nijhoff, Dordrecht – Boston – Lancaster 1987.
336. S. Sosnowski, J. Feng, M. A. Winnik, *J. Polym. Sci., A: Polym. Chem.* **32** (1994) 1497. Links
337. A. Malliaris, *Adv. Colloid Interface Sci.* **27** (1987) 153. Links
338. F. M. Winnik, S. T. A. Regismond, *Colloids Surfaces* **A118** (1996) 1.
339. G. Komaromy-Hiller, R. von Wandruszka, *J. Colloid Interface Sci.* **177** (1996) 156. Links

340. E. Alami, M. Almgren, W. Brown, J. Francois, *Macromolecules* **29** (1996) 2229. Links
341. D. Dibbern-Brunelli, T. D. Z. Atvars, *J. Appl. Polym. Sci.* **58** (1995) 779. Links
342. Ö. Pekcan, M. A. Winnik, M. D. Croucher, *Macromolecules* **23** (1990) 2673. Links
343. Ö. Pekcan, M. Canpolat, *J. Appl. Polym. Sci.* **59** (1996) 277. Links
344. M. Canpolat, Ö. Pekcan, *J. Appl. Polym. Sci.* **59** (1996) 1699. Links
345. M. Canpolat, Ö. Pekcan, *J. Polym. Sci., B: Polym. Phys.* **34** (1996) 691. Links
346. J. W. Haas III, M. M. Carrabba, R. W. Forney, *Field Screening Methods Hazard. Wastes Toxic, Chem Proc. Int. Symp.* (1995) 443.
347. R. Kotzick, R. Niessner, *Fresenius J. Anal. Chem.* **354** (1996) 72. Links
348. J. Lou, T. A. Hatton, P. E. Laibinis, *Anal. Chem.* **69** (1997) 1262. Links
349. S. Sasaki, Y. Ando, M. Dejima, Y. Arikawa, I. Karube, *Anal. Letts.* **31** (1998) 555.
350. Y. Wang, W. Liu, K. Wang, G. Shen, R. Yu, *Fresenius J. Anal. Chem.* **360** (1998) 702. Links
351. K. P. McNamara, X. Li, A. D. Stull, Z. Rosenzweig, *Anal. Chim. Acta* **361** (1998) 73. Links
352. J. E. Parks II, J. S. Armfield, T. E. Barber, J. M. E. Storey, E. A. Wachter, *Appl. Spectrosc.* **52** (1998) 112. Links
353. U. Panne, *Trends Anal. Chem.* **17** (1998) 491.
354. S. Bichlmeier, *Process* **12** (2005) 44.
355. P. Fichet, D. Menut, R. Brennetot, E. Vors, A. Rivoallan, *Appl. Opt.* **42** (2003) 6029.
356. C. K. Williamson, R. K. Daniel, K. L. McNesby, A. W. Miziolek, *Anal. Chem.* **70** (1998) 1186. Links
357. T. U. Wagner, T. Bundschuh, R. Köster, *Part. Part. Syst. Charact.* **22** (2005) 181.
358. J. M. Vadillo, P. L. Garcia, S. Palanco, D. Romero, J. M. Baena, J. J. Laserna, *Anal. Bioanal. Chem.* **375** (2003) 1144.
359. H. Fink, U. Panne, R. Niessner, *Anal. Chem.* **74** (2002) 4334. Links
360. G. Berden, R. Peeters, G. Meijer, *Int. Rev. Phys. Chem.* **19** (2000) 565. Links
361. M. H. Bharati, J. F. MacGregor, *Ind. Eng. Chem. Res.* **37** (1999) 4715.
362. B. Chu: *Laser Light Scattering*, Academic Press, San Diego 1991.
363. W. Brown: *Dynamic Light Scattering — The Method and Some Applications*, Oxford Science Publications, Oxford 1993.
364. W. Brown: *Light Scattering — Principles and Development*, Oxford Science Publications, Oxford 1996.
365. C. Kiparissides, J. F. MacGregor, S. Singh, A. E. Hamielec, *Can. J. Chem. Eng.* **58** (1980) 65. Links
366. T. Kourti, A. Penlidis, A. E. Hamielec, J. F. MacGregor, *Polym. Mat. Sci. Eng.* **53** (1985) 147.
367. T. Kourti, J. F. MacGregor, A. E. Hamielec, *Polym. Mat. Sci. Eng.* **62** (1990) 301.
368. J. A. Maroto, F. J. de las Nieves, *Colloid Polym. Sci.* **275** (1997) 1148. Links
369. J. A. Maroto, F. J. de las Nieves, *Colloids Surfaces* **A132** (1998) 153.
370. U. Apfel, R. Grunder, M. Ballauff, *Colloid Polym. Sci.* **272** (1994) 820. Links
371. U. Apfel, K. D. Hörner, M. Ballauff, *Langmuir* **11** (1995) 3401. Links
372. E. J. Röhm, K. D. Hörner, M. Ballauff, *Colloid Polym. Sci.* **274** (1996) 732.
373. A. O. Moser, O. Fromheim, F. Hermann, H. Versmold, *J. Phys. Chem.* **92** (1988) 6723.
374. E. Elbing, A. G. Parts, C. J. Lyons, B. A. W. Coller, I. R. Wilson, *Aust. J. Chem.* **42** (1989) 2085. Links
375. T. Kourti, J. F. MacGregor, A. E. Hamielec, D. F. Nicoli, V. B. Elings in C. D. Craver, Th. Provder (eds.): *Polymer Characterization*, Adv. Chem. Series. 227, American Chemical Society, Washington D.C. 1990
376. D. Horn, H. Auweter, W. Ditter, J. Eisenlauer, *Org. Coat.* 1986, 251.
377. D. Lilge, D. Horn, *Colloid Polym. Sci.* **269** (1991) 704. Links
378. S. Sadasivan, D. H. Rasmussen, *J. Colloid Interface Sci.* **193** (1997) 145. Links
379. H. Wiese, D. Horn, *Ber. Bunsenges. Phys. Chem.* **96** (1992) 1818.
380. P. C. Killgoar, R. A. Dickie, *J. Appl. Polym. Sci.* **21** (1977) 1813. Links
381. R. M. Fitch, R. C. Watson, *J. Colloid Interface Sci.* **68** (1979) 14. Links
382. A. R. Goodall, K. J. Randle, M. C. Wilkinson, *J. Colloid Interface Sci.* **75** (1980) 493. Links
383. R. M. Fitch, T. H. Palmgren, T. Aoyagi, A. Zuikov, *Angew. Makromol. Chem.* **123/124** (1984) 261.
384. F. B. Malihi, T. Provder, M. E. Koehler, *J. Coat. Technol.* **55** (1983) 41.
385. J. C. Thomas, V. Dimonie, *Appl. Optics* **29** (1990) 5332. Links
386. J. C. Thomas, *ACS Symp. Ser.* **472** (1991) 98.
387. G. Maret, P. E. Wolf, *Z. Phys. B — Condensed Matter* **65** (1987) 409. Links
388. S. Fraden, G. Maret, *Phys. Rev. Letts.* **65** (1990) 512.
389. D. Bicout, G. Maret, *Physica* **A 210** (1994) 87.
390. E. Sevick-Muraca, J. Pierce, H. Jiang, J. Kao, *AIChE J.* **43** (1997) 655.
391. M. Mohammadi, *Adv. Colloid Interface Sci.* **62** (1995) 17. Links
392. Z. Sun, S. Torrance, F. K. McNeil-Watson, E. M. Sevick-Muraca, *Anal. Chem.* **75** (2003) 1720. Links
393. F. H. Schwartz, M. Braun, US 5 900 933 (1999).
394. P. Barrett, *CEP Mag.* **99** (2003) 26.
395. U. Feller, B. Wessely, S. Ripperger, *Environm. Protect. Eng.* **25** (1999) 69.

396. M. Breitenstein, U. Kräuter, U. Riebel, *Part. Part. Syst. Charact.* **16** (1999) 249. Links
397. M. Breitenstein, U. Riebel, J. Shen, *Part. Part. Syst. Charact.* **18** (2001) 134. Links
398. X. Guo, U. Riebel, *Part. Part. Syst. Charact.* **22** (2005) 161. Links
399. F. J. Schorck in M. S. El-Aasser (ed.): *Advances in Emulsion Polymerization and Latex Technology*, 23rd Annual Short Course, Lehigh University, Bethlehem 1992.
400. J. Kim, S. Y. Jeong, K. U. Kim, Y. H. Ahn, R. P. Quirk, *J. Polym. Sci., A: Polym. Chem.* **34** (1996) 3277. Links
401. J. G. Dos Ramos, *Thesis,* Lehigh University, Bethlehem 1998.
402. J. Venkatesan, *Thesis,* Lehigh University, Bethlehem 1992.
403. C. M. Miller, E. D. Sudol, C. A. Silebi, M. S. El-Aasser, *J. Colloid Interface Sci.* **172** (1995) 249. Links
404. J. C. Giddings, *Science* **260** (1993) 1456. Links
405. J. C. Giddings, *Anal. Chem.* **67** (1995) 592A. Links
406. P. Blau, R. L. Zollars, *J. Colloid Interface Sci.* **183** (1996) 476. Links
407. L. Dulog, T. Schauer, *Progr. Coat.* **28** (1996) 25.
408. P. J. Wyatt, *J. Colloid Interface Sci.* **197** (1998) 9. Links
409. S. K. Ratanathanawongs, P. M. Shiundu, J. C. Giddings, *Colloids Surf. A* **105** (1995) 243.
410. D. Roessner, *Thesis,* University of Hamburg 1994.
411. P. J. Wyatt, *Anal. Chim. Acta* **272** (1993) 1. Links
412. R. Annino, R. Villalobos: *Process Gas Chromatography: Fundamentals and Applications*, Society of America, Research Triangle Park, NC 1992.
413. K. B. Beebe *et al.*, *Anal. Chem.* **65** (1993) 199R. Links
414. W. W. Blaser *et al.*, *Anal. Chem.* **67** (1995) 47R. Links
415. J. Workman *et al.*, *Anal. Chem.* **71** (1999) 121R. Links
416. V. Lopez-Avila, H. H. Hill, *Anal. Chem.* **69** (1997) 289R. Links
417. F. Diedrich, *Process* **10** (2003) 38.
418. R. Villalobos, R. Annino, *Adv. Instrum. Control* **46** (1991) 707.
419. R. G. Melcher, P. L. Morabito, *Anal. Chem.* **62** (1990) 2183. Links
420. W. R. Slater *et al.*, *Water Research* **24** (1990) 121. Links
421. E. R. Lindgren, D. W. Preshing, D. A. Kirchgessner, D. C. Drehmel, *J. Chromatogr.* **585** (1991) 353. Links
422. R. E. Berkley, *Process Control Qual.* **5** (1993) 159.
423. P. A. David, R. E. Pauls, *J. Chromatogr. Sci.* **36** (1998) 44. Links
424. C. J. Venkatramani, J. B. Phillips, *J. Microcolumn Sep.* **6** (1994) 229. Links
425. R. G. Mathews, J. Torres, R. D. Schwartz, *J. High Resol. Chromatogr. Chromatogr. Commun.* **1** (1978) 139.
426. M. G. Bloch, R. B. Callen, J. H. Stockinger, *J. Chromatogr. Sci.* **15** (1977) 504. Links
427. J. Clemmons, *Adv. Instrum. Control* **50** (1995) 97.
428. J. Durand, J. Beboulene, A. Ducrozet, *Analusis* **23** (1995) 481. Links
429. J. P. Durand, Y. Boscher, M. Dorbon, *J. Chromatogr.* **509** (1990) 47. Links
430. F.-Q. Ma, D.-S. Lu, Z.-Y. Guo, *J. Chromatgr. A* **740** (1996) 65.
431. K. Snavely, B. Subramaniam, *Ind. Eng. Chem. Res.* **36** (1997) 4413. Links
432. E. Dorrestijn, P. Mulder, *J. Anal. Appl. Pyrolysis* **44** (1998) 167.
433. V. Pacakova, V. Kozlik, *Chromatographia* **11** (1978) 266. Links
434. O. Kammona, E. G. Chatzi, C. Kiparissides, *J. Macromol. Sci. - Rev. Macromol. Chem. Phys.* **C39** (1999) 57.
435. A. Guyot, J. Guillot, C. Pichot, L. Rios Guerrero in D. R. Basset, A. E. Hamielec (eds.): Emulsion Polymers and Emulsion Polymerization, *ACS Symp. Ser.* **165** (1981) 415.
436. A. Urretabizkaia, J. R. Leiza, J. M. Asua, *AIChE Journal* **40** (1994) 1850.
437. M. Alonso, M. Recasens, L. Puigjaner, *Chem. Eng. Sci.* **41** (1986) 1039. Links
438. M. Alonso, M. Alivers, L. Puigjaner, M. Recasens, *Ind. Eng. Chem. Res.* **26** (1987) 65. Links
439. J. Janata, M. Josowicz, D. M. DeVaney, *Anal. Chem.* **66** (1994) 207R. Links
440. J. Janata, M. Josowicz, P. Vanysek, D. M. DeVaney, *Anal. Chem.* **70** (1998) 179R. Links
441. P. H. Müller, D. Werner, *Chemie Technik* **26** (1997) 48.
442. C. L. Guillemin, *Process Control Qual.* **3** (1992) 153.
443. R. E. Cooley, C. E. Stevenson, *Process Control Qual.* **2** (1992) 43.
444. N. C. van de Merbel, I. M. Kool, H. Lingeman, U. A. T. Brinkman, A. Kolhorn, L. C. de Rijke, *Chromatographia* **33** (1992) 525. Links
445. F. J. Sauter, *Am. Lab.* **10** (1992) 33.
446. E. Verette, F. Qian, F. Mangani, *J. Chromatogr. A* **705** (1995) 195. Links
447. Z. Ge, R. Thompson, D. Detora, T. Maher, P. Mckenzie, D. Ellison, J. Crutchfield, *J. Proc. Anal. Chem.* **3** (1997) 1.
448. M. G. Musolino, G. Neri, C. Milone, S. Minico, S. Galvagno, *J. Chromatogr. A* **818** (1998) 123. Links
449. T. Wessa, S. Küppers, G. Mann, M. Rapp, J. Reibel, *Org. Proc. Res. Dev.* **4** (2000) 102.
450. J. C. Thompson, R. H. Smith, *Proc. Control. Qual.* **2** (1992) 55.
451. L. Ebdon, H. W. Handley, P. Jones, N. W.BarnettJones, *Microchim. Acta* **2** (1991) 39.

452. M. A. Rey, J. M. Riviello, C. A. Pohl, *J. Chromatogr. A* **789** (1997) 149. Links
453. M. Nonomura, T. Hobo, *J. Chromatogr. A* **808** (1998) 151.
454. G. J. Lynch, *Proc. Control. Qual.* **1** (1991) 249.
455. R. M. Montgomery, R. Saari-Nordhaus, L. M. Nair, J. M. Anderson, *J. Chromatogr. A* **804** (1998) 55. Links
456. H. B. Wan, J. Liu, K. C. Ang, S. F. Y. Li, *Talanta* **45** (1998) 663. Links
457. B. A. P. Buscher, U. R. Tjaden, J. van der Greef, *J. Chromatogr. A* **788** (1997) 165. Links
458. B. E. Richter, B. A. Jones, N. L. Porter, *J. Chromatogr. Sci.* **36** (1998) 444. Links
459. Y. Shen, M. L. Lee, *J. Chromatogr. A* **778** (1997) 31. Links
460. R. Hoogendijk, *Anal. Chim. Acta* **378** (1999) 211. Links
461. M. H. Wright, A. M. James, *Kolloid-Z. Z. Polym.* **251** (1973) 745.
462. S. Bisal, P. K. Bhattacharya, S. P. Moulik, *J. Phys. Chem.* **94** (1990) 350.
463. L. F. J. Noël, R. Q. F. Janssen, W. J. M. van Well, A. M. van Herk, A. L. German, *J. Colloid Interface Sci.* **175** (1995) 461. Links
464. J. L. Reimers, F. J. Schork, *J. Appl. Polym. Sci.* **60** (1996) 251. Links
465. K. Voloudakis, P. Vrahliotis, E. G. Kastrinakis, S. G. Nychas, *Meas. Sci. Technol.* **10** (1999) 100. Links
466. A. F. Santos, E. L. Lima, J. C. Pinto, C. Graillat, T. McKenna, *J. Appl. Polym. Sci.* **90** (2003) 1213; Links A. F. Santos, E. L. Lima, J. C. Pinto, C. Graillat, T. McKenna *J. Appl. Polym. Sci.* **91** (2004) 941. Links
467. M. Kaminoyama, S. Taguchi, R. Misumi, K. Nishi, *Chem. Eng. Sci.* **60** (2005) 5513. Links
468. J. Q. Zhang, Y. Yan, *Powder Technol.* **135–136** (2003) 164.
469. A. M. Santos, P. Vindevoghel, C. Graillat, A. Guyot, J. Guillot, *J. Appl. Polym. Sci.* **34** (1996) 1271.
470. M. Patek, S. Bildstein, Z. Flegelova, *Tetrahedron Lett.* **39** (1998) 753. Links
471. S. Zoll, *Chemie Technik* **27** (1998) 34.
472. N. Le Bris, D. Birot, *Anal. Chim. Acta* **356** (1997) 205. Links
473. L. W. Niedrach, *Angew. Chem.* **99** (1987) 183. Links
474. T. Werner *et al.*, *Fresenius J. Anal. Chem.* **359** (1997) 150. Links
475. E. Pringsheim, E. Terpetschnig, O. S. Wolfbeis, *Anal. Chim. Acta* **357** (1997) 247. Links
476. U.-W. Grummt, A. Pron, M. Zagorska, S. Lefrant, *Anal. Chim. Acta* **357** (1997) 253. Links
477. D. A. Nivens, Y. Zhang, S. M. Angel, *Anal. Chim. Acta* **376** (1998) 235. Links
478. P. Westbroek, E. Temmerman, P. Kiens, *Anal. Comm.* **355** (1998) 21.
479. R. G. Cooks, S. H. Hoke, K. L. Morand, S. A. Lammert, *Int. J. Mass Spectrom. Ion Processes* **118** (1992) 1.
480. S. Bohatka, *Adv. Mass. Spectrom.* **13** (1995) 199.
481. M. A. DesJardin, S. J. Doherty, J. R. Gilbert, M. A. LaPack, J. Shao, *Proc. Control Qual.* **6** (1995) 219.
482. K. D. Cook, K. H. Bennet, M. L. Haddix, *Ind. Eng. Chem. Res.* **38** (1999) 1192. Links
483. R. Zimmermann, *Nachrichten Chem.* **50** (2002) 166.
484. R. Zimmermann, H. J. Heger, A. Kettrup, *Fresenius J. Anal. Chem.* **363** (1999) 720. Links
485. H. J. Heger *et al.*, *Anal. Chem.* **71** (1999) 46. Links
486. K. Hafner, R. Zimmermann, E. R. Rohwer, R. Dorfner, A. Kettrup, *Anal. Chem.* **73** (2001) 4171. Links
487. C. Didden, J. Duisings, *Proc. Control Qual.* **3** (1993) 263.
488. A. M. Gregory, L. A. Kephart, *Adv. Instrum. Control* **50** (1995) 933.
489. R. F. P. Nogueira *et al.*, *Ind. Eng. Chem. Res.* **38** (1999) 1754. Links
490. H. Deng, G. J. van Berkel, *Anal. Chem.* **71** (1999) 4284. Links
491. P. Fiorani, G. Margutti, G. Mariana, G. Moccia, *Proc. SPIE-Int. Soc. Opt. Eng.* **3509** (1998) 221.
492. J. J. Chambers, K. Min, G. N. Parsons, *J. Vac. Sci. Technol. B* **16** (1998) 2996. Links
493. V. M. Donelly, *Mater. Res. Soc. Symp. Proc.* **406** (1996) 3.
494. D. R. McKenzie, W. D. McFall, W. G. Sainty, Y. Yin, A. Durandet, R. W. Boswell, *Surf. Coat. Technol.* **82** (1996) 326.
495. S. Canegallo, G. Storti, M. Morbidelli, S. Carra, *J. Appl. Polym. Sci.* **47** (1993) 961. Links
496. F. Bleger, A. K. Murthy, F. Pla, E. W. Kaler, *Macromolecules* **27** (1994) 2559. Links
497. A. Penlidis, J. F. MacGregor, A. E. Hamielec, *J. Appl. Polym. Sci.* **35** (1988) 2023. Links
498. F. J. Schork, W. H. Ray, *J. Appl. Polym. Sci.* **28** (1983) 407. Links
499. F. J. Schork, W. H. Ray, *J. Appl. Polym. Sci.* **34** (1987) 1259. Links
500. R. G. Gilbert: *Emulsion Polymerization – A Mechanistic Approach*, Academic Press, San Diego 1995.
501. M. Holst, K. Schänzlin, M. Wenzel, J. Xu, D. Lellinger, I. Alig, *J. Polym. Sci, B: Polym. Phys.* **43** (2005) 2314. Links
502. W. R. Schowalter: *Mechanics of Non-Newtonian Fluids*, Pergamon, Oxford 1978.
503. H. A. Barnes, J. F. Hutton, K. Walters: *An Introduction to Rheology*, Elsevier, Amsterdam 1989.
504. C. W. Macosko: *Rheology: Principles, Measurements and Applications*, VCH Publishers, New York 1994.
505. K.-F. Arndt, G. Müller: *Polymercharakterisierung*, Carl Hanser Verlag, München – Wien 1996.
506. P. J. Halley, M. Mackay, *Polym. Eng. Sci.* **36** (1996) 593. Links
507. D. Hesekamp, H. C. Broecker, M. H. Pahl, *Chem. Ing. Tech.* **70** (1998) 286.
508. R. M. Fitch: *Polymer Colloids: A Comprehensive Introduction*, Academic Press, San Diego 1997.
509. E. J. Schaller in P. A. Lovell, M. S. El-Aasser (eds.): *Emulsion Polymerization and Emulsion Polymers*, John

Wiley & Sons, Chichester 1997.
510. R. Pal, *AIChE J.* **42** (1996) 3181.
511. R. Trim in *Polymer Testing '97 Design and Rheological Data in the Plastics Industry,* RAPRA, Paper 1, 1.
512. D. J. Fleming, T. W. Dobbie, A. L. Kelly, P. D. Coates in *Polymer Testing '97 Design and Rheological Data in the Plastics Industry,* RAPRA, Paper 2, 1.
513. A. Safranj, M. Yoshida, H. Omichi, R. Katakai, *Radiat. Phys. Chem.* **46** (1995) 987.
514. R. E. Challis *et al.*, *J. Appl. Polym. Sci.* **88** (2003) 1665. Links
515. S. A. Jones, H. J. Stronks, BRUKER minispec application note # 20.
516. Phillips, US 6 204 664, 2001 (M. Sardashti, X. Li).
517. S. J. Martin, V. J. McBrierty, D. C. Douglas, *Macromolecules* **34** (2001) 8934. Links
518. H. Fischer, O. Gyllenhaal, J. Vessman, K. Albert, *Anal. Chem.* **75** (2003) 622. Links
519. M. Maiwald, H. H. Fischer, Y.-K. Kim, H. Hasse, *Anal. Bioanal. Chem.* **375** (2003) 1111.
520. T. W. Skloss, A. J. Kim, J. F. Haw, *Anal. Chem.* **66** (1994) 536. Links
521. K. Landfester, S. Spiegel, R. Born, H. W. Spiess, *Colloid Polym. Sci.* **276** (1998) 356. Links
522. A. Nordon, C. A. McGill, D. Littlejohn, *Appl. Spectrosc.* **56** (2002) 75. Links

[Top of Page]

Abbreviations

AOTF: acousto-optic tunable filter

APACT: Advances in Process Analytics and Control Technology

APM: acoustic plate mode

ATR: attenuated total reflection

CCD: charge-coupled device

CHDF: capillary hydrodynamic fractionation

CID: charge-injection device

CPAC: Center for Process Analytical Chemistry

CPACT: Center for Process Analytics and Control Technology

CRDS: cavity ring-down spectroscopy

CTAC: Control Theory and Applications Center

CVD: chemical vapor deposition

DGEBA: diglicidyl ether of bisphenol A

DGEBF: diglicidyl ether of bisphenol F

DDM: diaminodiphenylmethane

DLS: dynamic light scattering

DWS: diffusing wave spectroscopy

EWA: evanescent wave absorption

FDA: Food and Drug Administration

FFF: field flow fractionation

FID: free induction decay

FIR: far infrared

FOCS: fiber-optic chemical sensor

FODLS: fiber-optic dynamic light scattering

FOQELS: fiber-optic quasi elastic light scattering

FRA: frequency response analyzer

FT: Fourier transform

FTIR: Fourier transform infrared

GC: gas chromatography

GDCh: Gesellschaft Deutscher Chemiker

HPCL: high perfomance liquid chromatography

HT: Hadamard transform

HDC: hydrodynamic chromatography

IC: integrated circuit

IFPAC: International Forum for Process Analytical Chemistry

IR: infrared

LEC: liquid exclusion chromatography

LED: light-emitting diode

LIBS: laser-induced breakdown spectroscopy

LIPS: laser-induced plasma spectroscopy

MACC: McMaster Advanced Control Consortium

MCEC: Measurement and Control Engineering Center

NIR: near infrared

NMR: nuclear magnetic resonance

PCS: photon correlation spectroscopy

PLS: partial least squares

PREACH: Process Related Environmental Analytical Chemistry

PS: power source

PVC: poly(vinyl chloride)

QELS: quasi-elastic light scattering

QCM: quartz crystal microbalance

RIM: reaction injection molding

SAW: surface acoustic wave

SEC: size exclusion chromatography

SERS: surface-enhanced Raman scattering

SH: shear-horizontal

SLS: static light scattering

SRS: stimulated Raman scattering

UV: ultraviolet

VIS: visible

XPS: X-ray photoelectron spectroscopy

Symbols

***A*:** absorbance (extinction), instrument constant

A_R**:** reactor wall heat exchange area

a η**:** Kuhn – Mark – Houwink exponent

***B*:** instrument constant

***c*:** concentration

C_0**:** equivalent capacitance of free space

c_i**:** concentration of component *i*

$c_{p,cool}$**:** specific heat capacity of coolant

$c_{p,k}$**:** specific heat capacity of *k*th reactor feed

$c_{p,liq}$**:** specific heat capacity of jacket liquid

$C_{p,mixture}$**:** heat capacity of reaction mixture

$C_{p,tot}$**:** heat capacity of filled reactor

***d*:** distance, path length

***D*:** diffusion coefficient, dielectric displacement

d_p**:** particle diameter

d_{stir}**:** stirrer diameter

***E*:** electric field

f_{res}**:** resonance frequency

***f*:** frequency, particle size distribution function

***g*:** contraction factor

***G*:** shear modulus, distribution function

g_1**:** electric field autocorrelation function

g_2**:** light intensity autocorrelation function

***G*:** conductance

***h*:** depth

$\Delta H_{r,j}$**:** reaction enthalpy of *j*th reaction

***I*:** electrical current, light intensity

I_0**:** incident light (radiation) intensity

I_d**:** detected light intensity

I_l**:** light intensity at distance *l*

***K*:** bulk modulus, scattering coefficient

K η**:** Kuhn – Mark – Houwink parameter

***l*:** (optical) path length, distance

k_Q**:** heat transfer coefficient

K_b: Boltzmann constant

m: refractive index ratio

m_k: mass of component k

m_{cool}: mass of coolant

m_{liq}: mass of jacket liquid

M_w: weight-average molecular mass

$M_{w,crit}$: critical weight-average molecular mass

n: refractive index

n_m: refractive index of medium

n_p: refractive index of particle

N_v: stirrer speed

p: pressure

P: polarization

p_0: pressure amplitude

p_{bubble}: bubble pressure

p_{gel}: gel point

p_h: hydrostatic pressure

p_x: pressure amplitude at distance x

P_0: power number

q: scattering vector

Q_{acc}: accumulated heat

Q_{cal}: heat generated by calibration heaters

Q_{el}: heat generated by electrical heaters

Q_{feed}: heat generated by material feed

Q_i: sources and sinks of heat

Q_{loss}: heat losses

Q_{osc}: heat of forced temperature oscillations

Q_r: heat of reaction

Q_{reflux}: heat flow over reflux condensor

Q_{stir}: heat input due to stirring

Q_{trans}: heat transfer through reactor wall

R: reflection coefficient, tube diameter

r: radius

r_h: hydrodynamic radius

R_j: reaction rate of *j*th reaction

t: time

T: transmission coefficient, temperature

δT: amplitude of temperature oscillations

T_1: outer reactor wall temperature

T_2: inner reactor wall temperature

$T_{cool,in}$: coolant temperature (inflow)

$T_{cool,out}$: coolant temperature (outflow)

T_g: glass transition temperature

T_J: reactor jacket temperature

$T_{J,in}$: jacket temperature (inflow)

$T_{J,out}$: jacket temperature (outflow)

T_k: temperature of *k*th reactor feed

T_{osc}: period of tube oscillations

$T_{osc,cal}$: period of tube oscillations filled with calibration liquid

T_R: reactor temperature

u: sound velocity

U: electrical voltage

u_0: sound velocity of reference component

U_0: initial electrical voltage

$(\Delta u / \Delta c)_i$: sound velocity-concentration coefficient *i*

$u_{l,liq}$: longitudinal sound velocity of liquids

$u_{l,sol}$: longitudinal sound velocity of solids

u_{mix}: sound velocity of a mixture

$u_{tr,liq}$: transversal sound velocity of liquids

$u_{tr,sol}$: transversal sound velocity of solids

U_x: electrical voltage at distance *x*

V: volume

V_{excess}: excess volume

V_i: volume of component *i*

V_r: volume of reaction mixture

V_{total}: total volume

***x*:** distance

x_r**:** monomer conversion

X_r**:** calorimetric conversion

***Y*:** complex admittance

***Z*:** acoustic impedance, complex impedance

Z'**:** real component of complex impedance

Z''**:** imaginary component of complex impedance

α**:** sound attenuation

α_{liq}**:** sound attenuation in liquids

β**:** adiabatic compressibility

δ**:** loss angle

ε**:** decadic extinction coefficient

ε^***:** dielectric permittivity

ε_0**:** free space dielectric permittivity

ε_∞**:** high frequency dielectric permittivity

ε'**:** storage component of dielectric permittivity

ε''**:** loss component of dielectric permittivity

$\varepsilon^*_{dipolar}$**:** dipolar reorientation contribution to ε^*

Υ**:** surface tension

Γ**:** decay constant

η**:** (bulk) viscosity

η_0**:** viscosity of dispersion medium

η_s**:** shear viscosity

η_v**:** volume viscosity

Φ**:** wave phase, volume fraction

Φ_r**:** phase shift between temperature oscillations

Λ**:** conductivity

λ**:** wave length

λ_{exc}**:** excitation wave length

λ_m**:** medium light wave length

λ_Q**:** heat conductivity

ν**:** frequency (of light), average velocity

θ**:** scattering angle

ϱ**:** density

ϱ_{cal}: density of calibration liquid

ϱ_{end}: density at reaction completion

ϱ_{mix}: density of reaction mixture

ϱ_{start}: density at reaction start

σ_{dc}: dc conductivity

τ: oscillation period, relaxation time, optical transmittance (turbidity)

ω: oscillation (angular) frequency

ω_{max}: frequency at loss maximum

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Pilot Plants

RICHARD P. PALLUZI, Exxon Research and Engineering Company, Annandale, NJ 08801, United States

1. **Introduction** 1
2. **Classification of Pilot Plants** 2
3. **Scale-Up** 2
4. **Planned Pilot Plant Experimentation** . 3
5. **Pilot Plant Design** 4
6. **Methods of Estimating Pilot Plant Costs** 5
7. **Pilot Plant Space** 6
8. **Scheduling** 6
9. **Pilot Plant Control System and Instrumentation** 7
10. **Feed and Product Handling** 8
11. **Pilot Plant Start-Up** 9
12. **Safety In Pilot Plant Design And Operation** 9
13. **References** 9

1. Introduction

While advances in modeling and predictive methods have greatly expanded the range and utility of theoretical research, there still can remain a significant degree of uncertainty in the final results before a major investment is justified. Small-scale laboratory tests help minimize this uncertainty but are fraught with risk due to the small scale and manual, batch simulation of most of the process steps. Hence the concept of piloting the process in a miniature unit in a way which more closely simulates the actual process plant. Pilot plants have evolved from multistory semi-works units, designed for one-tenth of final process scale, to microunits, which fit in larger laboratory hoods. Their cost and complexity often belie their small size and great utility.

The piloting process often occurs several times over the life of a research program, Table 1 shows a sequence of this piloting that might be part of a research program. Depending on the complexity of the process, the results of subsequent steps, and the organization's experience and comfort with the process and results, some of these steps may be eliminated. Recycle and feedback between each step is also possible when problems arise or concerns refuse to be resolved. As each step along this path involves an exponential increase in the resources, time, and money required, there is a strong incentive to progress to the next stage as soon as is practical. Conversely, there is a need to minimize the amount of recycling as the project progresses through the various steps to the commercial unit. While this rapid progress is desirable, the consequences of skipping a necessary preliminary stage or failing to fully understand the results of the previous stage can be devastating. Significant time and resources can be wasted at the next stage if it is focused along an incorrect or at least nonoptimum path. A substantial amount of time and money has progressively been invested, and any desire to change the process to make minor improvements must be resisted. At some point the decision must be made that the remaining unanswered process questions are acceptable risks.

Table 1. Typical research program steps

Step	Intention
Concept	Initiate ideas/concepts to be investigated
Basic research	Determine basic feasibility of concept
Preliminary economic evaluation	Determine if economic incentive is sufficient to proceed with investigation
Laboratory research	Develop basic data
Product evaluation	Evaluate product suitability
Process development/preliminary engineering	Produce preliminary design of actual commercial plant, resolving process questions as they arise
Pilot plant study	Prove basic reliability of proposed process or process design
Demonstration unit	Demonstrate process feasibility on small commercial scale
Prototype unit	Demonstrate process feasibility on small commercial scale
Commercial plant	Produce and sell a product at a profit

In today's climate of rapid change, timeliness can be as important or more important than minimizing the risks associated with a new product or process. Examples include securing a new mar-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.c19_c01

ket with a totally novel product or attempting to secure a patent position before a competitor. In this situation, the decision may be made to proceed to commercialization earlier than desirable, prior to satisfactorily resolving all design concerns. This often usually requires a more conservative (and more expensive) design approach, results in a significantly longer start-up, or produces a first unit plagued with operational difficulties. Resolving significant problems during start-up or in an operating unit, while feasible, is risky, expensive, and time-consuming. Rarely is it more effective than additional research.

A pilot plant is a collection of equipment designed and constructed to investigate some critical aspect(s) of a process operation or to perform basic research. It is a tool rather than an end in itself. A pilot plant can range in size from a laboratory bench-top unit to a facility approaching the size of a commercial unit. The purposes for its construction and operation can vary widely; they include:

- Confirming the feasibility of a proposed process
- Providing design data
- Determining the economic feasibility or a new process
- Determining optimum construction materials
- Testing the operability of a control scheme
- Determining the extent of plant maintenance
- Producing sufficient quantities of product for market evaluation
- Obtaining kinetic data
- Screening catalysts
- Proving areas of advanced technology
- Providing data for solutions to scale-up problems
- Providing technical support to an existing process or product
- Assessing process hazards
- Determining operating costs

The need for a pilot plant is a measure of the degree of uncertainty in developing a process from the research stage to a full commercial plant.

2. Classification of Pilot Plants

Pilot plants can be categorized by a number of different criteria. Size is the most common classification as it is the most uniformly proportional to construction and operating costs; a common classification is shown in Table 2. While a common classification, this approach shows little but a range of sizes and costs. The pilot plant's degree of automation is another classification as instrumentation typically represents large portion of the initial construction and annual operating costs; a common classification is shown in Table 3. While more useful in some respects, this classification is also primarily an indication of cost and complexity. In addition, the high cost of operating labor and the difficulties in manually a process while acquiring accurate data has forced the level of automation of virtually all but the simplest new pilot plants to be increased to such an extent that most fall into the highest automation class. Other classifications based on strategic purpose and/or design philosophy exist [1]. These are often more useful as shorthand methods of communicating useful concepts about pilot units in general. All these classifications are elastic.

3. Scale-Up

Scale-up is the process of developing a plant design from experimental data obtained from a unit orders of magnitude smaller (see also → Scale-Up in Chemical Engineering, Chap. 4.; → Process Development, Chap. 3.). This activity is considered successful if the commercial plant produces the product at design capacity, cost, and quality. This step from pilot plant to full-scale operation is the most potentially risky of all the phases of new process development, as the highest expenses are committed when the greatest risks occur.

Process plant design has come a long way from the early 1930s, when process designers were limited to rules-of-thumb that a process facility should not be scaled-up more than tenfold [2].

Table 2. Typical classification of pilot plants by size

Class	Characteristics	Cost range
Micro-unit or bench-top unit	Small size ($< 2\,m^2$), typically found in a hood or on a bench top Small tubing used throughput (3 – 6 mm diameter) Small feeds (Small size (< 4 L)	\$ 10 000 to \$ 100 000
Integrated pilot plant	Larger size (usually $10 - 25\,m^2$), located in open bay, walk in hood, or containment cell Tubing and small (13 – 25 mm diameter) pipe used Fully automated Feed and product systems with limited capacity usually included	\$ 50 000 to \$ 500 000
Demonstration or Prototype Unit	Very large size ($> 250\,m^2$), usually in a dedicated building or area Pipe used extensively Fully automated usually with a dedicated computer system Feed and product systems, including significant storage capacity included	\$ 1 000 000 and up

American Oil's Ultracracking unit at Texas City, for example, was designed with data from a small pilot plant with a scale-up factor of 80 000 [3]. However, many scale-ups still require smaller ratios, albeit greater than 10 : 1, particularly for new processes.

Scale-up problems exist for many reasons, including basic equations unsolvable by known mathematical techniques, interconnected physical and chemical process aspects resulting in coupled basic equations, solutions used on the pilot plant that are not suitable on the commercial scale, and unknown equipment performance at sizes never used before. Discussions of scale-up problems and solutions are available in [2–9]. While successful pilot plant operation does not guarantee successful commercial plant operation, it considerably decreases the risk. All stages of a research program should confirm good scale-up practices are being followed.

4. Planned Pilot Plant Experimentation

Planning the experimental program is an important step in determining the type of pilot plant required. This step should be performed as soon as the research program objectives are defined, as the type of experimental program required sets, to a large degree, the total program cost, the type of pilot plant required, and the required pilot plant operation. Hence planning the experimental program is one of the major effects on the economic justification of research programs. Planned experimentation

- Reduces the number of tests required – a key factor in controlling annual costs as it is not unusual for the annual operating costs of a pilot plant to be two to three times the initial construction costs.
- May alter the design or show that a pilot plant is not the best place to obtain some of the desired information.
- May highlight the need for more laboratory work to support the pilot plant operation. Laboratory work is less costly and frequently will give insight to correlations or models for the process to be studied in the pilot plant.
- Is instrumental in determining the overall project schedule.

Hence, the early inclusion of planned experimentation in any research program is vital. Pilot plant design specifications should be established only after careful experimental program development, as decisions on the accuracy of instruments, analyzers, and other equipment should be based on the planned experimental requirements, not haphazard assumptions. Flexibility and versatility are important but costly; when provided unnecessarily or too profusely, they can result in a unit that is expensive and time-consuming to construct and start up, as well as difficult or impossible to operate successfully. There is always a tendency to ensure the design covers any possible eventuality as well as to try to maximize the unit's use across future programs. Both tendencies often result in overly complex units, which often make achieving the original goal more difficult, costly, and time-consuming.

Statistical designs for experiments maximize information and reduce research time and costs. These techniques are less likely to miss synergistic factors affecting performance or product quality, minimize the element of human bias, eliminate less productive avenues of experimentation by taking advantage of previous data, and

reduce the number of pilot plant runs needed to define the effects of variables. Overall, statistical designs increase the confidence level in the experimental results. Additional information is available in [9–17], [70].

Table 3. Typical classification of pilot plants by degree of automation

Degree of automation	Characteristic	Cost of instrumentation ∗
Manual	]All process conditions set and monitored by operators Poor repeatability Loose process control	< $ 2000
Local control	Most important variables monitored by dedicated controllers Tighter process control Possibly some data-gathering instrumentation	$ 10 000 – $ 20 000
Mixed systems	All variables monitored by controllers (may be local or distributed) with some computer-controlled or -monitored Usually data gathering provided via "smart" data logger or computer Usually older system upgraded to more modern standards	$ 15 000 – $ 30 000
Automated	All or majority of systems computer-controlled Tightest process control May be supervisory or direct digital control High repeability Lowest operating labor Most batch operations may be automated Data gathering integral part of control system	$ 20 000 – $ 60 000

∗ Costs include installation of equipment but not field wiring and are based on a typical medium-sized pilot plant of 10 – 20 control loops. For smaller units (less than 10 control loops), reduce the above by 25 – 40 %; for larger units (20 – 40 control loops) increase costs by 50 – 100 %.

5. Pilot Plant Design

A pilot plant's design affects all aspects of its life and performance including its cost, operability, and effectiveness. While all operating ranges may not yet be fully defined at the initial design stage, a realistic preliminary range is required before the design is commenced, as is a clear definition of the pilot plant's purpose. The first step in pilot plant design is to determine whether to design a pilot plant for process modeling or problem investigation. Modeling the process involves reproducing the specific unit operations on a smaller scale. This promotes safe scale-up, minimizes design time, and reproduces all process operations of interest. It is usually expensive since all operations are reproduced, not just the most important ones. Investigating the problem involves designing a pilot plant to look at a specific area of interest. In this case, the pilot plant may not resemble the commercial operation. While this approach is usually cheaper and quicker, it carries the inherent risk of missing the real problem and producing nothing of value. Several references have addressed each type of approach [1], [18–21].

Pilot plants may be designed with conventional design techniques that mimic commercial process design. This generally provides for a safe and operable design, as the basic methodology is known and the effectiveness of the final results have been proven previously. While safer, this approach is not always feasible: the proposed scale of the pilot plant operation may be completely outside the range of all commercial design techniques, the design suggested may not be economical on the scale envisioned or in the location proposed, or copying the commercial design may carry some inherent limitations that adversely impact operations at ranges or conditions which are of interest in the pilot plant but not in the commercial plant. An alternative approach is to use a design methodology oriented to the pilot plant, by using many conventional design techniques, but trying to maximize the advantages of the pilot plant operation with regard to scale, techniques, and operation. Advantages include economic savings due to reduced equipment, construction, and operating costs; economical and efficient ways around otherwise expensive problems; increased simplicity of operation, resulting in reduced maintenance and operating staff, and improved process understanding through use of specialized equipment. The lack of published design information, however, raises the risk factor; hence, the quality of the final design is very dependent on the skill level and experiences of the design organization/engineer. A significant problem is the limited number of in-house pilot-plant-specific design organizations and truly pilot-plant-specific contractors. This makes using the alternative approach sometimes riskier, except for larger organizations with the requisite in-house resources

or after a diligent and often expensive contractor evaluation and qualification.

6. Methods of Estimating Pilot Plant Costs

Pilot plants costs range from US \$ 10 000 to 10×10^6, but most are commonly in the \$ 50 000 to \$ 250 000 range when an existing facility is available to house them. There are three basic methods for estimating the costs to design and construct a pilot plant; similarity, cost ratios, and detailed labor and materials. Similarity involves estimating the cost of the pilot plant on the basis of costs to design and construct a similar unit. While the fasted method, it is the least accurate, with errors of $\pm 100\,\%$ not uncommon. As pilot plants are generally built for new processes or process improvements, most are more or less unique. Few units are ever similar enough for good cost estimation. Cost ratios develop the cost estimate by relating the overall cost of the pilot plant or a part of the pilot plant to a known factor such as the cost of major process equipment, the number of control loops, the size of the equipment, or a variety of similar factors. The cost estimate is built up by using the ratios to develop the cost of the entire unit, individual equipment, or separate subsystems, depending on how detailed a cost ratio estimate is made. Unfortunately, cost ratio information, while widespread as a plant-estimating tool, is rarely available for pilot-plant-scale equipment. The lack of a large data pool also renders much of what is available suspect with regard to its overall accuracy. Depending on the type of cost ratios used and the experience of the personnel making the estimate, accuracies of $\pm 25-50\,\%$ are typical. Cost ratios can be used more accurately if they are restricted to small subsystems, as numerous information is available in this area [22]. This, however, usually requires a substantial increase in effort. Detailed labor and material estimation involves breaking the pilot plant construction down into a detailed series of small tasks and estimating the labor and materials required for each separate task. This estimation may be performed by in-house or contract engineers, obtained from contractors, or solicited, or solicited via bids or consulting. This method can produce estimates with accuracies of $\pm(10-20)\,\%$ but requires more effort than either of the previous methods. See Table 4 and [1], [22–28], [71] for further information.

In general, similarity or general cost ratios estimates are used for developing screening estimates when an estimate is required quickly and a lower accuracy is acceptable. Detailed cost ratios or detailed labor and materials estimates are generally developed prior to actual appropriation of funds so as to have a more accurate estimate for budgeting and cost control.

The costs for operating a pilot plant are a summary of the costs of the feedstock, product disposal, utilities, operating labor, spare parts, maintenance, and support services. Major efforts to reduce operating costs have focused on reducing the operating labor through automation and unattended operation [1], [29–37], as well as controlling waste-disposal costs by careful design and planning [1], [38].

While often overlooked or, at best, cursorily estimated in advance, operating costs outweigh construction costs within one to three years. Hence, investments in reducing these costs, while adding to initial unit costs, are generally prudent.

7. Pilot Plant Space

Pilot plant space can be divided into several basic types: separate buildings, containment cells or barricades, open bays, walk-in hoods, and laboratory areas. Space is often selected more for availability than suitability or long-term operability. This can lead to significant problems later. For a more detailed summary of the advantages and disadvantages of each type of space and a full discussion of the impacts, see [1]. The space required for a pilot plant will vary tremendously with its size and type. A small unit may require only part of a laboratory (perhaps $5-10\,m^2$) while an "average" pilot plant may require a large room or building (perhaps $500-2000\,m^2$), excluding extended feed or product storage.

Pilot plant space often requires significant exhaust ventilation, large numbers of utilities, and special power and climate control. Hence, it is expensive to construct and even more expensive to operate.

Table 4. Pilot plant cost estimating techniques

Method	Accuracy range	Information required	Time to develop ∗ (weeks)
Similarity	± (50 – 100) %	Costs of similar unit Differences from proposed unit	< 1
General cost ratios	± (25 – 50) %	All scaling elements (such as control loops, size or similar factors) Different from ratio basis	1 – 2
Detailed cost ratios ∗∗	± (15 – 30) %	Detailed P & ID Complete equipment list Preliminary layout	4 – 12
Detailed labor and materials	± (10 – 20) %	Detailed P & ID Preliminary electrical and instrumentation drawings Complete equipment and materials list Preliminary layout	4 – 12

∗ After preliminary design is completed. ∗∗ The difference between a general and a detailed cost ratio cost estimate is the level of detail involved. A general cost ratio estimate might develop the total cost for the pilot plant control system – or even the pilot plant – from the number of control loops. A detailed cost ratio estimate would estimate the cost per loop based on the type of loop. The rest of the estimate would have to be generated by a large number of similar ratios.

8. Scheduling

A major concern in all pilot plant work is minimizing the time involved between project inception and meaningful data generation. Typically, an "average" pilot plant requires six or eighteen months to progress through this process, as shown on Table 5. This lead-time can be reduced significantly through careful planning, purchasing long-lead-time items early and various construction techniques. Reference [1] gives typical lead times for major pieces of equipment for planning purposes. Design time can be reduced if standardized designs have been developed in advance for common subsystems. For a discussion of problems with crash pilot plant projects, see [69].

Table 5. Typical lead times

Task	Time, months
Design, materials, procurement, scheduling	3 – 12
Construction	3 – 12
Start-up (up to and including first actual run)	1 – 6

9. Pilot Plant Control System and Instrumentation

Defining the requirements for a pilot plant control system is often difficult, as data on research-scale systems is rarely available, often contradictory, and usually less than convincing. Using the more readily available process control system data often results in a system that is too large, too complex, and more expensive than is really required for a pilot plant. The plethora of small PC-based control and data-acquisition systems on the market makes evaluation of all possible alternatives a tedious and time-consuming task. The evaluation process is not straightforward and requires experience and skills not always readily available to a research organization. Personnel inexperienced in either instrumentation systems or pilot plant operations frequently perform the control system design. The isolated and often intermittent nature of pilot plant operations also inhibits evolution and promotes individual unique installations. This complicates the selection process.

Automatic instrumentation systems for pilot plant were originally based on small stand-alone analog devices. The trend in these devices has been towards smaller and less expensive units with more flexibility; this increases their suitability for pilot plant instrumentation systems. Computer-based control systems, once restricted to large process plant applications, are now commonly available in sizes and costs suitable for most pilot plant operations. The additional data gathering and work-up capabilities of these systems make them so attractive that they have largely displaced analog systems for all but the simplest applications. Computer-based systems continue to grow smaller and less expensive. Analog equipment is also being redesigned as small stand-alone digital units. This is blur-

ring the distinction between computer-based and stand-alone systems. Most organizations are installing computer-based systems on all but the simplest pilot plant.

Three types of computer control systems are commonly used for pilot plant control systems. The first is a centralized system, historically based on a minicomputer or occasionally a mainframe but now increasingly network based. These systems have large storage capacities, substantial memories, and numerous associated equipment. They typically control all the pilot plants in an area or facility. Centralized systems are economical if a large number of units are involved but are less common due to their high installation and maintenance costs, as well as the limitation that any failure of the central system shuts down all the pilot plants involved. The growth of in-house networked computing systems has reversed this decline and resulted in a greatly increased growth in this area. The resulting systems work like stand-alone PCs but have the advantages of central systems. In effect they are starting to become the low cost DCS systems (see below) pilot plant organizations have been searching for.

Stand-alone computer systems, usually based on a personal computer (PC) or programmable logic controller (PLC), provide a separate system for each pilot plant. This allows for economical expansion for new units, separates pilot plants completely for maintenance and troubleshooting, and often has the lowest initial cost. Standardization can be a problem, and software control, data gathering, and storage packages can be limited in size, scope, and capability; these are usually acceptable trade offs.

Distributed control systems (DCS) are a hybrid of the previous systems, with stand-alone computers for control and data gathering at each pilot plant, tied together with one or more higher level computers for data storage and work-up. Typically more expensive than either central systems or stand-alone computers, distributed control systems have the advantages of both systems. This approach is preferred for pilot plants when it can be economically justified, usually for larger research facilities or very large demonstration units. As mentioned above, networked, stand-alone PCs are beginning to lower the overall costs to the extent that many organizations will be able to take advantage.

For further information on the different systems, including a more detailed discussion of their costs, advantages, and disadvantages, see [1], [40]; for information on actual experience, see [32], [34–36], [41–47], [72], [73].

The support resources necessary to service, repair, and calibrate the control system must also be considered when selecting a system; they are considerable and can be estimated from the literature [48–50].

Pilot plants are usually heavily instrumented compared to commercial plants. For further information on some of the more common pilot plant instrumentation see [1], [51], [74].

Analytical instrumentation for pilot plants are divided into two general classes, off-line and on-line. Off-line analysis is a batch operation requiring sample taking, handling, and storage followed by analysis later in time, frequently in another location. It interferes least with the process, is commonly less expensive, and allows the analysis to be performed under rigorously controlled conditions. (Some analyses cannot be made in any other manner due to equipment limitations, although this is becoming less common). Disadvantages include the difficulty in taking, handling, and storing representative samples and the inability to modify process on a real-time basis due to the time lag between sampling and analysis.

On-line analysis is often more expensive and difficult to set up initially but can be more accurate and reliable if performed properly on-line monitoring of Chemical Reactions. On-line analyzers can also be used to provide real-time control of a secondary process variable such as severity or conversion, as opposed to controlling a primary variable, such as temperature [36], [52]. The growth in this area suggests that fewer and fewer analyses will be done off-line, and that these will be restricted to difficult or capital intensive tests [75].

The maintenance of analytical instrumentation requires trained personnel and is time-consuming [39], [40]. An additional problem is the necessity of frequently checking the calibration of the analysis instrumentation and recalibrating if required.

Stand-alone data-gathering instrumentation, once common in pilot plants, has been virtually replaced in all but the simplest pilot plant by a

data-gathering computer, usually used for process control as well.

10. Feed and Product Handling

One of the most vexing aspects of pilot plant work can be feed and product handling, as a pilot plant is neither designed nor operated as a closed loop system like a commercial plant. Feedstock is often obtained by different methods with different purity. Products are often not separated or processed as completely. The problems involved in handling and storing feed and product materials can often rival the pilot plant process problems in difficulty.

The toxicity of both feeds and products must be carefully considered during the preliminary design stages [53], [54], especially if they contain known or suspected carncinogens. Special testing may be required prior to operation to determine the toxicity of unknown materials.

Provision must be made for sufficient storage of both feeds and products while minimizing operator-handling time. Sufficient surge time must be available in both feed and product systems to minimize the need for contained close operator attention. High level alarms and cutoffs should be installed on all product vessels that could overflow; low-level alarms and cutoffs should be considered for all feed vessels that could be emptied. Vented products must conform to all existing federal, state, and local environmental regulation, as well as good engineering judgement. This may require scrubbing or neutralization, incineration, flaring or similar measures. Care is required to ensure emergency venting is adequately dimensioned, selected, and installed and that products are vented to an appropriate height and/or location. The area electrical classification around the unit must be carefully evaluated and selected. If a pilot plant is located in an electrically classified area, handling of flammable feeds and products becomes more difficult. The same is true if toxic materials are used in a poorly ventilated area. For a further discussion of some of these concerns, see [68].

11. Pilot Plant Start-Up

Pilot plant start-up is different from major process plant start-up due to the smaller scale of he unit, smaller resources committed, lack of advance start-up planning, and limited experience with the pilot plant process and operation. Yet is it often as complex and demanding a task. The resultant underestimating, limited planning and staffing often creates problems.

The key a successful pilot plant start-up is advance planing; at least 6 – 8 weeks before project completion to allow identification of problem areas and concerns in time for successful resolution. A detailed start-up plan should be developed, listing each task to be performed in chronological order. The start-up plan then allows a list of required resources and a tentative start-up schedule to be developed. A successful start-up requires this advance planning as well as adequate trained and experienced personnel with a variety of skills. Many companies have found that a specialized start-up group is a major asset if pilot plant work is regularly done. Safety is also a major concern during start-up, as interlock systems are not fully functional and equipment and subsystems are being energized for the first time. For a more detailed discussion of pilot plant start-up including developing a start-up plan, see [1].

Pilot plant start-up costs vary widely; the costs booked to the start-up as well as the extent of the commissioning activities labeled as "start-up" varies widely among organizations. Typical costs range from 5 to 50 % of construction cost, with 10 to 20 % typical. Start-up durations also vary greatly, but a range of one to three months is common, depending on the personnel training and expertise. One often-overlooked anomaly is the increasing time required for computer programming during start-up. This is due to the large increase in automation for most units. Despite "user-friendly" systems, programming still often requires special training and skills not easily or economically developed in operating personnel. Another problem is skill retention, as trained operating personnel typically lose their edge when called upon to use programming skills only very intermittently. Hence many organizations are being forced to provide specialists who routinely provide this service to a

larger base. When unrecognized or unavailable, this can lead to significant start-up delays.

12. Safety In Pilot Plant Design And Operation

Pilot plants are often more hazardous than process plants, even though they are smaller in size. This is due to a tendency to relax standard safety review procedures (based on he small scale, exceptionally qualified personnel involved, and the experimental nature of the research operations), the lack of established operational practice and experience, the lack of information regarding new materials or processes, and the lack of effective automatic interlocks (due to the frequently changing nature of pilot plant operations, the desire for wide latitude in operating conditions, and the lack of full-time maintenance personnel). To minimize these concerns, most organizations require a formal series of safety reviews and hazard analysis for each new or modified pilot plant. At a minimum, this involves analyzing the proposed pilot plant before construction to identify and eliminate any potential hazards. This includes toxicity or flammability concerns, feed or product handling, disposal problems, relevant regulations, and potentially harmful reactions. The exact type and timing of these reviews varies between organizations; see [1], [39], [40], [68] for a further discussion.

The pilot plant must also be carefully designed so that its control and safety systems are fail-safe and any unexpected equipment or utility failure brings the unit into a safe and de-energized condition. Unexpected or rapid process changes, if they can herald or lead to dangerous conditions, such as runaway exothermic reaction, should be continuously monitored by appropriate instrumentation and suitable automatic action provided.

Further information can be found in [1], [55–68].

13. References

1. R. P. Palluzi: *Pilot Plants: Design, Construction and Operation,* McGraw Hill, New York 1992.
2. R. Katzen, *Chem. Eng. (NY)* **75** (1968) March, 95.
3. F. R. Bradbury, *Chem. Tech.* **3** (1973) 532.
4. H. C. Henry, J. B. Gilbert, *Ind. Chem. Process Des. Develop.* **12** (1973) March, 328.
5. J. M. Smith, *Chem. Eng. Prog.* **64** (1968) Aug., 78.
6. P. Trambouze, *Chem. Eng. (NY)* **86** (1979) Sept.
7. P. Trambouze, *Chem. Eng. (NY)* **97** (1990) Feb.
8. R. H. Wang, L. T. Fan, *Chem. Eng. (NY)* **81** (1974) May.
9. L. Gore, *Ind. Eng. Chem.* **43** (1965) Oct., 2327.
10. A. C. Ackinson, *Chem. Eng. (NY)* **73** (1966) May, 149.
11. W. G. Hunter, A. C. Ackinson, *Chem. Eng.* **159** (1966) June.
12. K. A. Brownlee: *Statistical Theory and Methodology in Science and Engineering,* John Wiley & Sons, New York 1965.
13. W. G. Cochran, D. R. Cox: *Experimental Designs,* John Wiley & Sons, New York 1965.
14. W. E. Biles, J. J. Swain: *Optimization and Industrial Experimentation,* John Wiley & Sons, New York 1980.
15. O. L. Davies: *The Design and Analysis of Industrial Experiments,* Longman, New York 1978.
16. O. L. Davies, O. L. Goldsmith: *Statistical Methods in Research With Special Reference to the Chemical Industry,* Longman, New York 1976.
17. C. R. Hicks: *Fundamental Concepts in the Design of Experiments,* Holt, Reinhart and Winston, New York 1973.
18. L. A. Robbins, *Chem. Eng. Prog.* **75** (1979) Sept., 45.
19. J. B. Maerker, J. W. Schall, *Chem. Eng. Prog.* **45** (1953) Aug., 1622.
20. D. E. Garret, *Chem. Eng. Prog.* **55** (1959) Sept., 44.
21. J. W. Payne, *Ind. Eng. Chem.* **45** (1953) Aug., 1621.
22. Numerous volumes on cost data have been published by R. S. Means Company Some examples include: *Means Mechanical Cost Data, Building Construction Data and Means Electrical Cost Data,* R. S. Means Company, Kingston, MA.
23. R. Perry et al.: *Perry's Chemical Engineers' Handbook,* 6th ed., McGraw Hill, New York 1984.

24. J. W. Drew, A. F. Ginder, *Chem. Eng. (NY)* **77** (1970) Feb.
25. C. Shanmugam, *Hydrocarbon Process.* (1983) Jan.
26. D. S. Remer, L. H. Chai, *Chem. Eng. Prog.* **86** (1990) Aug.
27. D. S. Remer, L. H. Chai, *Chem. Eng. Prog.* **86** (1990) April.
28. K. R. Cressman, *Cost Eng.* **25** (1983) Feb., 31.
29. C. E. Ellis, *Chem. Eng. Prog.* **64** (1968) Oct., 50.
30. J. W. Carr, Jr., et al., *Chem. Eng. Prog.* **61** (1965) June, 84.
31. C. D. Ackerman, *Chem. Eng. Prog.* **61** (1965) June, 67.
32. C. D. Ackerman, A. B. Hartmann, R. A. Wright, *Ind. Eng. Chem. Process Des. Develop.* **6** (1967) April, 476.
33. L. C. Hardison, G. P. Huling, K. J. Metzger, *Chem. Eng. Prog.* **67** (1971) April, 50.
34. W. M. Herring, S. E. Shields, *Chem. Eng. Prog.* **61** (1965) June, 94.
35. L. C. Hardison, R. A. Lengemann, *Chem. Eng. Prog.* **60** (1964) June.
36. R. D. Dean, N. B. Angelo, R. B. Young, *Chem. Eng. Prog.* **61** (1965) June.
37. C. E. Ellis, *Chem. Eng. Prog.* **64** (1968) Oct.
38. M. M. Girvotch, J. G. Lowenstein, *Chem. Eng. (NY)* **97** (1990).
39. R. P. Palluzi, *Chem. Eng. (NY)* **97** (1990) March, 76.
40. R. P. Palluzi, *Chem. Eng. Prog.* **87** (1991) Jan., 21.
41. C. D. Ackerman, P. Hurling, K. J. Metzger, *Chem. Eng. Prog.* **67** (1971) April.
42. E. J. Bourne, W. M. Herring, *Chem. Eng. Prog.* **73** (1977) Dec.
43. M. K. Corrigan, *AIChE Annual Meeting,* San Francisco, CA, Nov., 1989.
44. W. E. Kaufman, *Chem. Eng. Prog.* (1979) May.
45. L. Uitenham, R. Munjal, *Chem. Eng. Prog.* **87** (1991) Jan.
46. R. K. Wang, L. T. Fan, R. A. Van Sweringen, *Chem. Eng. Prog.* **53** (1961) Nov., 129.
47. P. Tessier, N. Therien, *Microprocessors Microsys.* **11** (1987) March, 99.
48. A. T. Upfold, *Instrument. Tech.* **18** (1971) Feb., 46.
49. F. H. Barbin, *Instrument. Tech.* **20** (1973) Nov., 44.
50. R. W. Purgh, Paper 13 D, AIChE National Meeting, Tulsa, OK, March 10 – 13, 1974.
51. B. Liptak (ed.): *Instrument Engineers' Handbook,* Chilton, Radnor, PA 1982.
52. J. E. Osborne, *Chem. Eng. Prog.* **70** (1974) Nov., 76.
53. C. H. Powell, *Chem. Eng. Prog.* **67** (1971) April, 71.
54. F. A. Graf, Jr., *Chem. Eng. Prog.* **63** (1967) Nov., 67.
55. R. N. Brummel, *Plant Operations Prog.* **8** (1989) Oct., 228.
56. M. A. Caparo, J. Strickland, *Chem. Eng. Prog.* **85** (1989) Oct., 13.
57. A. R. Jones, D. J. Hopkins, *Chem. Eng. Prog.* **62** (1966) Dec.
58. R. W. Prugh, *Chem. Eng. Prog.* **63** (1967) Nov., 49.
59. V. Lesins, J. J. Moritz, *Chem. Eng. Prog.* **87** (1991) Jan.
60. J. C. Dore, *Plant Operations Prog.* **7** (1988) Oct., 223.
61. J. W. Carr, Jr., *Chem. Eng. Prog.* **84** (1988) Sept., 52.
62. P. C. Ashbrook, M. M. Renfrew: *Safe Laboratories,* Lewis, Chelsa, MI 1991.
63. Center for Chemical Process Safety (ed.): *Guidelines For Hazard Evaluation Procedures,* American Institute of Chemical Engineers, New York 1992.
64. F. A. Graf, *Chem. Eng. Prog.* **63** (1967) Nov.
65. H. T. Kohlbrand, *Chem. Eng. Prog.* **81** (1985) April.
66. V. Siminski, *Chem. Eng. Prog.* **83** (1987) Jan.
67. L. V. Beckman, *Chem. Eng. (NY)* **99** (1992) Jan.
68. R. P. Palluzi: *Pilot Plant and Laboratory Safety,* McGraw Hill, New York 1992.
69. R. P. Palluzi, *CEP Chem. Eng. Prog.* **93** (1997) Dec.
70. K. K. Hochman, D. Berengut, *Chem. Eng.* (1995) Nov.
71. K. Popham, *AACE Trans* (1996).
72. J. LaFuci, *Chem. Eng. Prog.* **94** (1998) March.
73. P. J. Cody, A. P. Claire, *Pharm. Eng.* (1995) July/Aug.
74. N. E. Battikiha, *InTech.* **45** (1997) May.
75. D. E. Padkulski, *Chem. Eng. Prog.* **93** (1997) Oct.

Pimelic Acid → Dicarboxylic Acids, Aliphatic
Pinacolone → Ketones

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Polymerase Chain Reaction

Standard Article

Hermann Willems[1], Cornelie Jäger[1], Gerald Reiner[2]

[1]Institute for Hygiene and Infectious Diseases of Animals, Justus-Liebig-University, Giessen, Germany

[2]Department of Swine Diseases, Justus-Liebig-University, Giessen, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.c21_c01.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Abbreviations and Glossary of Special Terms
2. Introduction
3. Principle
4. Reaction Components
4.1. DNA Polymerases
4.1.1. Fidelity
4.1.2. Mg^{2+} as Cofactor
4.1.3. Inhibitors
4.2. Deoxynucleoside Triphosphates (dNTPs)
4.3. Reaction Buffer
4.4. Primer
4.4.1. Melting Temperature
4.4.2. Concentration
4.4.3. Synthesis
4.4.4. Primer Design
4.5. Template
5. Equipment and Laboratory Preconditions
6. Cycling Conditions and Reaction Phases
6.1. Denaturation Conditions
6.2. Annealing Conditions
6.3. Extension Conditions
6.4. Amplification Phases
7. Characterization of PCR Products
7.1. Gel Electrophoresis
7.2. Restriction Analysis
7.3. Hybridization
7.4. Sequencing
8. Modifications of PCR for Increased Specificity and Sensitivity
8.1. Increasing Specificity
8.2. Increasing Sensitivity
9. Contamination
9.1. Sources of Contamination
9.2. PCR Controls
9.3. Decontamination Methods
9.3.1. Uracil-N-Glycosylase (UNG)
9.3.2. UV Light

9.3.3. Enzymes
9.3.4. Psoralens and Isopsoralens
10. Quantitative PCR
10.1. External Control
10.2. Internal Control
10.3. Real-Time Quantification
10.3.1. Fluorescence Resonance Energy Transfer (FRET)
10.3.2. Förster Energy Transfer (TaqMan System)
11. Applications
11.1. Detection of Pathogenic Agents
11.1.1. Preconditions
11.1.2. Extraction of Template DNA
11.1.2.1. Isolation of Total DNA
11.1.2.2. Isolation of Pathogen Specific DNA
11.1.3. Amplification
11.1.4. Multiplex PCR
11.1.5. In Situ PCR
11.1.6. Universal PCR
11.1.7. PCR Controls for Diagnostic Purposes
11.2. Detection of Genetic Disorders or Cancerous Cells
11.2.1. Known Mutations
11.2.1.1. Amplification Refractory Mutation System (ARMS)
11.2.1.2. Allele-Specific Oligonucleotides (ASOs)
11.2.1.3. Introduction of Restriction Sites
11.2.1.4. Competitive PCR
11.2.1.5. Primer Extension PCR (PEST)
11.2.2. Unknown Mutations
11.2.2.1. Denaturing Gradient Gel Electrophoresis (DGGE)
11.2.2.2. Single Strand Conformation Polymorphism (SSCP)
11.2.2.3. Chemical Cleavage of Mismatches (CCM)
11.3. PCR Analysis in Evolution and Taxonomy
11.3.1. Restriction Fragment Length Polymorphism
11.3.2. PCR for Typing Satellites or HLA Genes
11.3.3. Random Amplified Polymorphic DNA
11.3.4. Analysis of Fingerprints
11.3.5. Restrictions and Future Aspects
11.4. Tissue Typing
11.5. PCR in Food Analysis
11.6. PCR to Engineer DNA
11.6.1. Cloning of PCR Products
11.6.2. PCR Mutagenesis
11.7. PCR for Analysis of Gene Expression

[Top of Page]

1. Abbreviations and Glossary of Special Terms

Allele:	=	one of two or more variants of a gene
Amplicon:	=	PCR-generated specific DNA fragments of defined length
Amplification:	=	multiplication of DNA or RNA fragments by PCR
Annealing:	=	hybridization of oligonucleotides to complementary sequences of DNA
AP-PCR:	=	arbitrarily primed PCR, synonymous for RAPD
ARMS:	=	amplification refractory mutation system
ASO:	=	allele-specific oligonucleotide
bp:	=	base pair; complementary nucleotides that form characteristic bonds
Blunt end:	=	end of double-stranded nucleic acid molecule where both strands have the same length
Booster PCR:	=	PCR technique for increased specificity; reaction is started with low primer concentrations and primer concentration is increased during PCR

CCD: = charge-coupled device
CCM: = chemical cleavage of mismatches
cDNA: = complementary DNA copy of a mRNA molecule synthesized by reverse transcriptase
Complementarity: = specific binding of adenine to thymine (or uracil) and guanine to cytosine between two antiparallel strands in nucleic acids molecules
DAF: = DNA amplification fingerprinting; PCR with short random primers generating a set of unspecific amplicons
Denaturation: = melting of double-stranded nucleic acids in single-stranded molecules by temperature or chemical agents
DGGE: = denaturing gradient gel electrophoresis
Dideoxynucleotide: = analogon to deoxynucleotide without 3′-hydoxy-group; synthesizing of DNA molecules with such nucleotides terminates elongation
dNTP: = deoxynucleotide
dsDNA: = double-stranded DNA
ERIC-PCR: = enterobacterial repetitive intergenic consensus PCR; fingerprinting technique for characterization of microbes
Exonuclease: = nucleic acids restricting enzyme; restriction starts at the 3′- or 5′ end of the molecule
FRET: = fluorescence resonance energy transfer
GMO: = genetically modified organism
HLA: = human leukocyte antigen
Homology: = identical DNA or RNA sequence in two different molecules
Hot start PCR: = PCR with two compartments initially separated by wax; amplification is started at elevated temperatures when the wax layer melts and the two compartments are mixed
Hybridization: = linking of two single stranded nucleic acid molecules by complementary base pairs
IMS: = immunomagnetic separation
LED: = light emitting diode
MHC: = major histocompatibility complex
MUX: = multiplexer
Nested PCR: = PCR technique for increased sensitivity and specificity; two consecutive PCR reactions are performed; the amplicon of the first reaction serves as template in the second PCR
PEST: = primer extension PCR
Plasmid: = circular DNA molecule which carries genetic elements permitting its own autonomous replication in bacteria and yeast; sometimes coding for characteristics of microbes like resistance to antibiotics
Primer: = oligonucleotide which anneals specifically to complementary DNA; primer serve as starter molecules for the elongation of complementary DNA strands by polymerases
Probe: = DNA, RNA or oligonucleotide that binds (hybridizes) to complementary single stranded DNA or RNA
Processivity: = number of nucleotides incorporated by a polymerase before the enzyme drops off the DNA matrix
Proofreading: = capacity of polymerases to recognize and substitute incorrect bases while synthesizing new nucleic acid molecules
RAPD: = random amplified polymorphic DNA; applied for differentiation of species or subspecies
REP-PCR: = repetitive extragenic palindromic PCR; fingerprinting method
Restriction enzyme: = enzymes restricting DNA at specific palindromic sequences
Reverse transcriptase: = enzyme that synthesizes single stranded DNA complementary to RNA matrix
RFLP: = restriction fragment length polymorphism
rRNA: = ribosomal RNA
RT-PCR: = reverse transcription followed by PCR
Satellite DNA: = repetitive DNA elements in eucaryotes; number of repetitions varies individually
SSCP: = single strand conformation polymorphism
ssDNA: = single-stranded DNA
Sticky end: = double-stranded nucleic acid molecules with 3′ or 5′ protruding ends generated by restriction enzymes
Stringency: = conditions (temperatures, salt concentrations) influencing specificity of PCR
Template: = matrix DNA for the synthesis of new DNA molecules by polymerases during PCR
Touch down PCR: = PCR with particular temperature conditions for increased specificity
UNG: = uracil-N-glycosylase
UPGMA: = unweighted pair group method, averages
VNTR: = variable number of tandem repeats
XL PCR: = extra long PCR for amplification of large DNA fragments

[Top of Page]

2. Introduction

Polymerase chain reaction (PCR) technology enables scientists to rapidly generate large amounts of genetic material from trace amounts which would otherwise be too small to analyze. Or in other words: PCR is a method to amplify minute amounts of particular DNA fragments (in the most extreme case one molecule) from a complex mixture of nucleic acids (DNA, RNA) so as to analyze the fragment or to make it visible by other techniques. The method has been discovered by K. B. MULLIS in 1983 [16], [17] and has revolutionized molecular biology. Since that time numerous applications of the PCR method were developed, reflected in the exponential increase in the number of publications. In 1993 K. B. MULLIS eventually received half of the Nobel Prize for Chemistry due to his epoch-making invention. PCR is simple to perform, fast, flexible and combines high specificity and high sensitivity in contrast to other detection methods where specificity and sensitivity are always incompatible. Due to its simplicity it has soon been introduced into many other fields, such as forensic science, diagnosis of genetic diseases, evolutionary biology, environmental science, food analysis, and so on.

[Top of Page]

3. Principle

The PCR reaction consists of three steps (equivalent to one cycle) which are repeated 25 to 35 times. During the first step template DNA is denatured (Figs. 1-3, step 1). Subsequently, two oligonucleotides (primers) anneal to complementary sequences of the single-stranded DNA (ssDNA) (Figs. 1-3, step 2). Primer sequences are then elongated by a thermostable DNA polymerase to recover double-stranded DNA (dsDNA) (Figs. 1-3, step 3). The products generated during the first and second cycle are of undefined length. Beginning with the third cycle amplicons with a defined length flanked by the two primers are produced (Figs. 1-3, white boxes). In the following cycles the number of defined amplicons increases exponentially whereas products of undefined length accumulate in a linear mode.

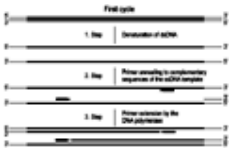

Figure 1. PCR principles — first cycle

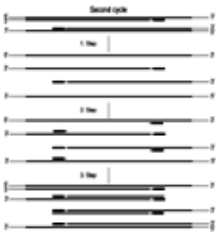

Figure 2. PCR principles — second cycle

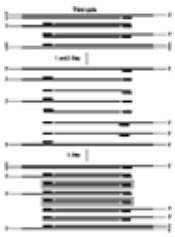

Figure 3. PCR principles — third cycle

The PCR temperature profile is shown in Figure 4. Denaturing of dsDNA is usually performed at 93 – 97 °C (step 1) for 15 to 60 s. Primer annealing temperature depends on the physicochemical properties of the oligonucleotides and ranges from 50 – 70 °C (step 2). Extension of primer sequences (step 3) by a thermostable DNA polymerase occurs at 72 °C for a few seconds to several minutes dependent on the length of the DNA fragment to be amplified. Because of the high temperatures, thermostability of polymerases is crucial for PCR reaction. In case of insufficient thermostability, enzyme (e.g., Klenow fragment ⟶ Enzymes) must be added for each cycle [18].

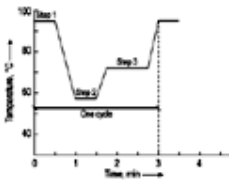

Figure 4. PCR temperature profile

The PCR reaction mixture contains reaction buffer (Section Reaction Buffer), two different primers (Section Primer), and a

thermostable DNA polymerase (Section DNA Polymerases) which needs magnesium chloride as cofactor and deoxynucleotides (dNTPs) (Section Deoxynucleoside Triphosphates (dNTPs)) as substrates. DNA or RNA may serve as template. RNA needs to be reverse transcribed to cDNA prior to PCR amplification.

Standard PCR. Because of the great number of parameters influencing the product yield it is impossible to specify conditions applicable to each PCR reaction. Nevertheless, parameters outlined below may be a starting point for optimizing PCR conditions.

The reaction volume may be chosen between 20 and 100 µL consisting of reaction buffer, primers (each 1 µM), DNA polymerase (0.2 to 1 U, i.e., International Units, ⟶ Enzymes – Definition of Units), $MgCl_2$ (1.5 mM) and dNTPs (100 – 200 µM each). A temperature profile of 94 °C for 30 s, 55 °C for 30 s and 72 °C for 1 min works in most cases. Usually 35 cycles are enough to visualize the PCR product on ethidium bromide stained agarose gels under UV illumination.

[Top of Page]

4. Reaction Components

4.1. DNA Polymerases (see also ⟶ Nucleic Acids – DNA Replication)

Since the introduction of PCR the number of thermostable DNA polymerases [19] has been increasing continuously. *Taq* DNA polymerase, the prototype of thermostable DNA polymerases had been isolated from the bacterium *Thermus aquaticus* in 1968 by BROCK [20]. Since that time many other thermostable DNA polymerases have been isolated (Table 1). Thermostable DNA polymerases can be differentiated by their nuclease activity, thermal stability, optimal reaction conditions, and their reading accuracy (fidelity). DNA polymerases having a 5′-3′ nuclease activity are capable of repairing DNA which has been damaged (e.g., by UV light, X-rays, or chemicals). Polymerases with a 3′-5′ nuclease activity are proofreading enzymes which recognize wrongly inserted nucleotides and replace them by right ones. Polymerases having both functions are those with the greatest fidelity and thus with the lowest error rate.

The concentration of the DNA polymerase varies between 1 and 5 U per 100 µL PCR reaction.

Table 1. Properties of various thermostable DNA polymerases

DNA polymerase	Source	Exonuclease activity		Half-life at 95 °C, min
		3′-5′	5′-3′	
Pwo	*Pyrococcus woesii*	yes	no	?
Taq	*Thermus aquaticus*	no	yes	40
Pfu	*Pyrococcus furiosus*	yes	yes	120
Vent	*Thermococcus litorales*	yes	no	400
Deep Vent	*Thermococcus litorales*	yes	no	1380
Tfl	*Thermus flavus*	no	no	?
Tth	*Thermus thermophilus*	yes	no	20
Ultma	*Thermotoga maritima*	yes	no	50

Application of a recombinant thermostable DNA polymerase combined with modified buffer composition and cycling conditions allow amplification of DNA fragments of up to 40 000 base pairs (bp) [21], [22]. Amplification of long fragments is designated as XL (extra long) PCR and the most widely used enzyme for this purpose is r*Tth* polymerase (recombinant *Thermus thermophilus* polymerase) [23], [24]. This polymerase requires magnesium acetate instead of $MgCl_2$ and the optimal extension temperature is 68 °C. Usually annealing and extension are carried out at the same temperature (68 °C). Template quality is decisive for successful XL PCR. Therefore, template DNA must be handled very carefully to avoid disruption.

4.1.1. *Fidelity*

The proportion of altered DNA molecules after PCR amplification *f* depends on the DNA polymerase error rate per cycle *p*, the number of cycles *n*, and the number of target molecules prior to PCR amplification s_0:

$$f = \frac{(s_0 \cdot p)(n \cdot 2^{n-1})}{s_0 \cdot 2^n} = \frac{n \cdot p}{2}$$

with $s_0 \cdot p$ being the error per cycle, and $n \cdot 2^{n-1}$ the increase in DNA molecules during the PCR process. For example, the error frequency of an enzyme with an error rate of 1×10^{-4} amounts to 0.1 % after 20 cycles of amplification, which equals one error per 1000 nucleotides. Under the same conditions a polymerase with an error rate of 5×10^{-6} produces only five errors in 200 000 nucleotides which is one error per 40 000 nucleotides. Accordingly, error frequency can be reduced simply by decreasing the number of cycles and/or using a polymerase with both, 5′-3′ and 3′-5′ exonuclease activity, i.e., a high fidelity. This is especially important when PCR products are to be cloned.

4.1.2. *Mg^{2+} as Cofactor*

Mg^{2+} is a cofactor of almost all DNA polymerases with an optimal concentration between 1.0 and 5.0 mM. A high Mg^{2+} concentration causes formation of unspecific products and a low amount of Mg^{2+} ions results in little or no PCR product. The Mg^{2+} concentration needs to be optimized for each PCR reaction.

4.1.3. *Inhibitors*

Numerous substances (Table 2) are known to inhibit processivity of DNA polymerases [25]. Such inhibitors are often contained in sample materials and may be copurified with template DNA.

However, it has also been demonstrated that the sensitivity of various DNA polymerases to inhibitors present in samples is different [26]. For example, when r*Tth* was used instead of *Taq* DNA polymerase it became possible to amplify DNA in the presence of 20 vol % blood without reduced amplification sensitivity, whereas *Taq* DNA polymerase was totally inhibited in the presence of 0.004 vol % blood in the PCR mixture.

Most frequently DNA is isolated from blood which contains EDTA or heparin to prevent coagulation. Both substances inhibit DNA polymerases. Therefore, it is important that blood samples are not preserved with coagulation inhibitors. Porphyrin compounds in the blood itself are also known to prevent amplification. This is the reason why cells (leukocytes, bacteria, yeast, etc.) must be separated from erythrocytes to eliminate porphyrin compounds. Meanwhile, DNA extraction kits are commercially available (e.g., Qiagen, BioRad) which are not only to be used for DNA extraction from blood samples but also from tissue, bacteria, viruses, etc.

DNA prepared by phenol/chloroform extraction must be free of traces of phenol and sodium dodecyl sulfate (SDS) since both inhibit DNA polymerases. Phenol can be removed by ethanol precipitation of DNA. Many DNA extraction methods are based on proteinase K treatment to lyse cells or bacteria. Since proteinase K hydrolyzes proteins the enzyme must be removed or inactivated because otherwise it would destroy the DNA polymerase. Heating of the suspension for 10 min to 100 °C inactivates proteinase K.

Table 2. PCR inhibitors

Substance	Concentration
Sodium docecyl sulfate (SDS)	> 0.005 wt %
Phenol	> 0.2 vol %
Ethanol	> 1 vol %
Isopropanol	> 1 vol %
Sodium acetate	≥ 5 mM
Sodium chloride	≥ 25 mM
EDTA	≥ 0.5 mM
Hemoglobin	≥ 1 mg/mL
Heparin	≥ 0.15 U/mL
Urea	> 20 mM

Samples being suspected of containing inhibitors should be treated in one of the following ways:

1. *Dilution:*
   Dilution of samples decreases the inhibitor concentration in the sample and thus enables amplification. Because dilution of samples is coupled with a loss in PCR sensitivity this technique should be used only when a sufficient amount of DNA in the samples is to be expected.
2. *Precultivation of bacterial samples:*
   Since PCR amplifies not only DNA from living but also from dead bacteria the message of PCR results has to be evaluated. If only living bacteria are to be detected a precultivation step is indispensable. By this way inhibitors are diluted and the amount of template DNA is increased. Thus PCR sensitivity is also increased, but the additional step involves the danger of contamination.
3. *Nested PCR (see also Section Increasing Sensitivity):*
   If there is some remaining activity of the DNA polymerase after the PCR or if new polymerase is added, two consecutive PCRs may be performed. Products from the first round of PCR are reamplified in a second PCR with primers directed against an internal region of the PCR product. Since only a small volume of the first PCR reaction is added to the second PCR, inhibitors are diluted.
4. *DNA extraction:*
   In most cases inhibitors are removed by DNA extraction. These methods are described in Section Extraction of Template DNA.

4.2. Deoxynucleoside Triphosphates (dNTPs)

Deoxynucleotides (dATP, dCTP, dGTP, dTTP) are used in final concentrations (f.c.) of 50 – 200 µM each. Higher concentrations may cause errors of the DNA polymerase. If lyophilized powder of dNTPs is to be used, the solution must be neutralized to avoid degradation of the nucleotides. It is more convenient to use ready-to-use dNTP mixes which are commercially available from many suppliers (e.g., Carl Roth, Biozym, Applied Biosystems, MWG Biotech). One point

should be kept in mind when varying dNTP concentration: Mg^{2+} is bound quantitatively by dNTPs, therefore the amount of Mg^{2+} ions has to be adjusted whenever changing the dNTP concentration.

4.3. Reaction Buffer

In general, PCR buffer consists of 10 mM *N*-Tris(hydroxymethyl)aminomethane (Tris), pH 8.4, 50 mM KCl, 0.01 % gelatin, and nonionic detergents. Nonionic detergents (NP-40, Tween 20, both in a concentration of 0.1 vol %) are needed to achieve maximum processivity of the DNA polymerase. Reaction buffer as well as $MgCl_2$ are supplied together with the DNA polymerase by the manufacturer.

4.4. Primer

4.4.1. *Melting Temperature*

According to definition the melting temperature T_m is the temperature where 50 % of the DNA molecules are double-stranded and 50 % are single-stranded. The T_m depends on the guanosine – cytosine (G – C) content of the DNA, the salt and the primer concentration. If the G – C content of the primers is nearly 50 % T_m may be approximated by the following equation:

$$T_m = 2 \cdot (\text{no. A and T}) + 4 \cdot (\text{no. G and C})$$

For example, the oligonucleotide 5′-AGC TGC CTA GAT CGA GTA CGT TGC-3′ is composed of 11 As and Ts and 13 Gs and Cs, thus the T_m can be calculated as:

$$T_m = 2 \cdot 11 + 4 \cdot 13 = 64^\circ\text{C}.$$

The melting temperature (T_m) is calculated automatically if using commercially available primer design software packages.

4.4.2. *Concentration*

In most cases primers are delivered as lyophilized powder. Although the enclosed primer data sheet includes the mass of the primers it is good practice to determine the concentration of primers after dissolving in an appropriate solvent (e.g., bidistilled water or 10 mM Tris/1 mM EDTA). The primer concentration of this solution can be calculated from the optical density (OD) at 260 nm. For primers with a G – C content around 50 % the following equation can be applied [27]:

$$c_{\text{Primer}} = \text{OD}_{260} \cdot 33[\text{ng}/\mu\text{L}] \text{ or}$$

$$c_{\text{Primer}} = \text{OD}_{260} \cdot 100/n[\text{pmol}/\mu\text{L}]$$

where *n* is the number of bases.

4.4.3. *Synthesis*

Oligonucleotides are synthesized fully automatically on a solid phase support. The solid phase consists either of controlled pore glass (CPG) or polystyrene particles, which are already loaded with the first base to be synthesized. Consequently, there are A, G, C, and T columns. Synthesis starts at the 3′ end of the primer. During synthesis modified bases may be incorporated, e.g., fluorophores, biotin or digoxigenin labeled bases, inosine, etc. If primers are to be constructed from protein sequences inosine is the base of choice for the wobble positions in a codon. Inosine may be incorporated by polymerases into growing dsDNA opposite each of the four bases (A, G, C, or T). Such oligonucleotides are designated as “degenerated”. DNA synthesis services are offered from many companies (e.g., MWG Biotech, Carl Roth, Applied Biosystems) in different scales (0.01 – 10 µmol).

4.4.4. *Primer Design*

A precondition for successful establishment of a PCR is the choice of convenient primers. Many primer design software programs are now available which allow to set appropriate options and to seek for compatible primer pairs automatically. Basically, there are four different ways to obtain primer sequences or sequences from which primers are to be deduced:

1. From literature
2. From database entries
3. From newly obtained sequences in the course of a sequencing project, and
4. From protein sequences

Basic conditions for the primers to be fulfilled are: high specificity and efficient annealing to the sequence of interest. This is especially important if there is only a minute amount of DNA to be amplified out of a high background of foreign DNA. The following parameters should be taken into consideration to design compatible primers:

1. A balanced base distribution
2. No stretches (more than 4) of bases and/or no repetitive sequence motifs
3. No stable secondary structures (self complementarities) especially at the 3′ end
4. No complementary sequences between primers (primer dimer formation), especially at the 3′ end
5. 15 – 30 bases in size
6. Final concentration 0.05 – 1 µM
7. Unique sequence
8. A maximum of 1 – 2 Gs and/or Cs at the 3′ end

9. T_m of both primers should be similar (± 2 °C), and
10. The size of the PCR product should be in the range of 100 – 600 bp.

The length of the primers can vary between 15 and 30 bases though most of them have sizes of 20 ± 2 bases. The standard concentration of primers is 1 µM. Only in special cases, e.g., when unspecific products are produced, specificity can be increased by decreasing the primer concentration.

Software. Primer design software may be ordered, e.g., from the following suppliers:

1. Medprobe AS, Oslo, Norway,
2. Molecular Biology Insights, Cascade, USA,
3. Scientific & Educational Software, Cary, USA,
4. Premier Biosoft International, Palo Alto, USA.

There are numerous freeware software programs available online which may be downloaded from the following URLs:

http://www.biocenter.helsinki.fi/bi/Programs/download.htm (FastPCR)

http://www.rsbs.anu.edu.au/Products&Services/BiotechnologyTransferUnit/gprime.asp (Gprime)

http://perlprimer.sourceforge.net/download.html (Perlprimer)

http://www.charite.de/bioinf/tgge/ (TGGE-Star)

Furthermore, primers can also be designed online under the following URL:

Autoprime:	http://www.autoprime.de/
Amplify:	http://bip.weizmann.ac.il/mb/bioguide/pcr/PCRsofAmplify.html
Codeshop:	http://blocks.fhcrc.org/blocks/codehop.html
GeneFisher:	http://bibiserv.techfak.uni-bielefeld.de/genefisher/
Methprimer:	http://www.urogene.org/methprimer/index1.html
Oligomaster:	http://www.oligomaster.co.uk/Downloads (limited to 10 oligos)
Primer 3:	http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi
PrimerX:	http://bioinformatics.org/primerx/
PrimerQuest:	http://biotools.idtdna.com/primerquest/
Primo:	http://www.changbioscience.com/primo/
PDA:	http://dbb.nhri.org.tw/primer/
Pride:	http://www.dkfz-heidelberg.de/tbi_old/services/Pride/search_primer
Oligo2002:	http://www.bioinformatics.vg/bioinformatics_tools/oligo2002.shtml
Webprimer:	http://seq.yeastgenome.org/cgi-bin/web-primer

Primo is a package of programs allowing to design primers for many different applications, e.g., Standard PCR, Multiplex PCR, degenerate PCR, gene synthesis, RAPD.

4.5. Template

PCR may be performed with dsDNA, ssDNA, or RNA as template. The DNA fragment which is to be amplified must be deduced from a genetically stable region to obtain consistent PCR results. If possible, abundant DNA regions like repetitive DNA elements, rDNA regions, multicopy genes, or mitochondrial DNA should be chosen. In this way the sensitivity of the PCR reaction can be increased, since one template molecule contains several copies of the target DNA. This is especially true for PCR reactions where minute amounts of template DNA are expected.

If the total length of the genome *L* in base pairs and the DNA concentration is known, the number of molecules *n* in the PCR reaction prior to amplification can be calculated:

$$n = \frac{m \cdot 10^{12}}{1.1 \cdot L}$$

with *m* being the DNA amount in nanograms.

The PCR detection limit depends strongly on the material from which DNA has to be amplified. In the most extreme case the detection limit is one target molecule per PCR reaction. As a rule of thumb the volume of an unknown DNA-containing sample (e.g., clinical samples) should be no more than 1/10 of the total PCR reaction volume.

[Top of Page]

5. Equipment and Laboratory Preconditions

Reaction tubes (*PCR cups*) for DNA amplification are usually characterized by small size and thin walls for fast

temperature transduction. Exact pipetting of smallest amounts (0.2 – 1 µL) is crucial for PCR. Thus *micropipettors* for PCR must be of highest reliability.

Thermocycler. Soon after its discovery the PCR process was automated by using two approaches: robotics and temperature cycling devices. Most current cyclers are of the latter type. In robotics the temperature control stations are stationary and the samples are moved between these stations (Stratagene, robocycler), whereas in the temperature cycling devices the samples remain in the same physical position while sample block temperature is cycled under software control.

Due to the enormous progress in durability of Peltier elements almost all cycler types are based on this technique. Nevertheless, differences in ramp times for heating and cooling make it difficult to transfer cycle parameters from one model to another model of cycler. Mostly the cycle parameters need to be adapted. A heatable cover lid is a must for currently employed cyclers making it possible to omit overlaying the PCR reaction with mineral oil.

If the cycler is to be used in routine diagnostics with a high throughput of samples the cycler should have a high capacity (96, 384 wells) and accept microtiter plates. Cycler with several temperature blocks are advantageous if PCRs with different cycle parameters are to be used at the same time or if PCR parameters need to be optimized. Table 3 lists thermal cyclers with their corresponding features.

Table 3. Thermal cycler models

Supplier	Name	Capacity, cups per block	Special features	Real-time detection
Stratagene	Robocycler	40/96	Robotics	no
Biometra	-TGradient	48/96/384	gradient capability, slide cycling option	no
	Tpersonal	20/48		no
	T1	48/96/348	slide cycling option	no
Bio-Rad	MJ Mini Cycler	48	gradient capability	no
	PTC 100	60/96		no
	DNA Engine family	30/48/60/96/384	swappable alpha-units (blocks), slide cycling options, optional motorized lid	optional (Chromo 4 alpha unit), 4 channels
	Opticon 2	96	gradient capability	yes, 2 channels
	MiniOpticon	48	gradient capability	yes, 2 channels
	iQ5	96	gradient capability	yes, 5 channels
	MyiQ	96	gradient capability	yes, 1 channel
Eppendorf	Mastercycler			
	personal	25/16	1 universal block for 0.2 and 0.5 mL tubes, personal card	no
	gradient	96/77	1 universal block for 0.2 and 0.5 mL tubes, personal card, gradient capability	no
	Realplex 2, Realplex 4	96	gradient capability	yes, 2/4 channels
Applied Biosystems	2720	96		no
	GeneAmp 9700	60/96/384	interchangeable blocks	no
	7300	96		yes, 4 channels
	7500	96	fast version available (PCR in less than 40 min)	yes, 5 channels
	7900HT	96/384	PCR in less than 40 min, continuous wavelength detection from 500 – 660 nm	yes
Roche Diagnostics	Lightcycler			
	480	96/384	PCR in less than 40 min (384 block) and 1 h (96 block),	yes, 6 channels

			respectively	
	1.5/2.0	32	PCR in 20 min	yes, 3/6 channels
LTF-Labortechnik	Rotor-Gene	32/72/100	open system adaptable to all real-time techniques and dyes, spinning of samples at 500 rpm therefore no temperature variation from sample to sample	yes, 6 channels

PCR in a Diagnostic Laboratory. To perform PCR, including sample preparation, analysis and documentation at least three separate rooms or compartments should be available.

Room no. 1 is used to prepare PCR reagents and the PCR reaction mix (mastermix without DNA). For pipetting, weighing and measuring of chemicals only unused, non-autoclaved plastic ware should be used. The usage of these chemicals should be restricted to this single purpose.

In room no. 2 samples are received, prepared, and PCR reactions loaded with sample DNA. It is convenient to prepare samples in a laminar flow work bench with an integrated UV light source. All preparation steps should be carried out with sealed pipette tips to avoid contamination of the pipettor shaft with template DNA.

In room no. 3 amplification, electrophoresis, and documentation of PCR results take place.

There should be no direct connection between rooms and no air circulation from post-PCR to pre-PCR rooms. Access to these rooms should be restricted. Equipment should remain in the respective rooms and not moved between rooms. For each room separate protective clothing should exist and gloves have to be changed regularly.

[Top of Page]

6. Cycling Conditions and Reaction Phases

6.1. Denaturation Conditions

Complete denaturation of template DNA is necessary to achieve a high amplification efficiency. Typical parameters for denaturation are 93 to 95 °C for 15 to 30 s. If the template DNA has a high G – C content (> 60 %) denaturing time should be prolonged to 1 min or a pre-PCR denaturation step of 5 min should be performed. Denaturation time depends also on the PCR cups which are used. In general, only thin-walled PCR cups should be used to minimize temperature transition time from the heating block to the interior of the PCR cup.

6.2. Annealing Conditions

Annealing temperature and time depend on the base composition, the size and concentration of primers (Section Primer). These parameters have to be optimized for each PCR to achieve a high PCR product yield without unspecific side products.

6.3. Extension Conditions

Normally, the optimal extension temperature for the DNA polymerase is known. Activity of most thermostable polymerases is optimal at about 72 °C. Extension time depends on the size and concentration of the amplicon and is different for each DNA polymerase. For example, *Taq* DNA polymerase has a temperature optimum of 72 °C and a processivity at this temperature of 35 to 100 nucleotides per second. Thus, an extension time of 1 min would be enough to synthesize a DNA fragment of 2000 bases in size. However, since the amount of DNA increases with time and after a few cycles exceeds enzyme concentration (about 1 nM) extension time has to be adapted. A starting point for optimization extension time should be 10 s per 100 base pairs.

6.4. Amplification Phases

From a certain cycle on exponential amplification of the DNA template switches to a linear and finally to a stationary phase. The stationary phase is designated as “plateau”. The plateau effect can have the following reasons:

1. Consumption of substrates (dNTPs and/or primers),
2. Inactivation of the DNA polymerases and/or the dNTPs,
3. Inhibition of the DNA polymerase by excess of dsDNA and pyrophosphates,
4. Competition by unspecific products or primer dimers,
5. Reassociation of specific products at concentrations above 10^{-8} M, and
6. Incomplete denaturation of the specific product at high product concentrations

The plateau effect can be avoided by decreasing the DNA amount before amplification and/or by reduction of the cycle numbers. In most cases the amount of PCR product is sufficient for downstream manipulations (gel electrophoresis Section Gel Electrophoresis, restriction analysis Section Restriction Analysis, sequencing Section Sequencing, etc.).

[Top of Page]

7. Characterization of PCR Products

7.1. Gel Electrophoresis

Electrophoresis of DNA is performed on agarose or polyacrylamide gels. In electric fields DNA molecules migrate to the positive electrode due to their negatively charged surface. Migration of molecules depends only on the size if the molecules are equally charged. Hence, length of DNA fragments may be determined by gel electrophoresis.

7.2. Restriction Analysis

Endonucleases type II restrict double-stranded DNA at specific restriction sites consisting of characteristic base sequences so that sticky or blunt ends of DNA are generated. Restriction of amplicons results in DNA fragments of characteristic and previously determined length if the DNA sequence of template DNA is known (Fig. 5).

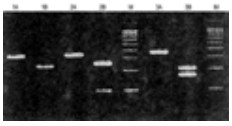

Figure 5. Restriction analysis of different PCR products (1, 2, and 3). Amplicons in lanes 1A, 2A, and 3A have been restricted with *Not*I generating two fragments in each case (1B, 2B, and 3B).

M = molecular mass marker

7.3. Hybridization

For identification of DNA fragments, i.e., amplicons, hybridization techniques are considered to be very reliable tools. During hybridization double-stranded DNA is generated from complementary strands with different origin. The sequence of one strand is known and thus binding allows conclusions to be drawn about the sequence of the other strand. The molecule with known sequence is labeled and designated as probe. Prior to hybridization amplified DNA (eventually after electrophoresis) is transferred to membranes where hybridization is carried out [28].

7.4. Sequencing

Nowadays DNA sequencing is mainly carried out automatically based on cycle sequencing principles [29] with labeled dideoxynucleotides as chain-terminating component [30], [31].

[Top of Page]

8. Modifications of PCR for Increased Specificity and Sensitivity

8.1. Increasing Specificity

Apart from optimizing $MgCl_2$, polymerase (Section Mg^{2+} as Cofactor), or primer concentration (Section Concentration) further techniques exist for increasing specificity of PCR.

If very small amounts of DNA (i.e., 100 templates or less) are contained in the sample, in the first few cycles nucleotides, primers, and DNA polymerase are present in a great excess and therefore the probability that unspecific products or primer dimers are formed is very high. In this case it is advisable to change the PCR technique and use techniques like Booster, Hot Start, and Touch Down PCR.

With currently available thermocyclers it is possible to check annealing temperatures for maximum specificity within one PCR run. This is achieved by thermocycler blocks allowing to generate a temperature gradient over the block cavities (Thermal cycler with gradient capability, Table 3). Such an approach is especially important for institutions which must establish a great number of different PCR methods or which must perform different PCR assays with primer pairs having incompatible annealing temperatures.

Booster PCR. In Booster PCR the primer concentration at the very beginning of amplification is very low and additional primers are added after a few rounds of amplification. However, this technique is used very rarely since the PCR cup has to be opened for a second time and thus the risk of contamination increases.

Hot Start PCR. In Hot Start PCR two compartments are formed which are separated by a wax layer. The lower compartment contains PCR buffer, $MgCl_2$, nucleotides, and primers and the upper compartment contains PCR buffer, DNA polymerase, and template DNA. When the cycling process starts the wax layer is liquefied, the two compartments are merged and amplification is initialized.

In the late 1990s DNA polymerases which do not exhibit any activity at room temperature were introduced (Applied Biosystems, Qiagen). Inactivity at ambient temperatures is achieved by a chemical modification of DNA polymerases or blocking of the enzyme with an antibody [32]. Activity is fully released only when the system is heated to 95 °C for several minutes. This kind of Hot Start PCR eliminates the need of separating PCR reagents by a wax layer, it is performed like a “normal” PCR reaction. As mentioned above, full activity of the DNA polymerase in early cycles of PCR is not necessary. Therefore, the activity of these enzymes may be only partially released by decreasing the retention time at 95 °C. In subsequent cycles the remaining activity is released stepwise and thus the half-life of the polymerase can be increased so that PCRs with up to 60 cycles can be performed. Thereby, not only specificity but also sensitivity of the PCR reaction is improved [33].

Touch Down PCR. In Touch Down PCR the first 5 to 10 cycles are performed at annealing temperatures which are very stringent so that unspecific primer annealing is prevented. For example, in the first cycle primer annealing is performed at 67 °C and subsequently annealing temperature is decreased by steps of 2 °C per cycle until the optimal annealing temperature of 57 °C is reached. The remaining cycles are performed at 57 °C.

8.2. Increasing Sensitivity

Although PCR is characterized in general by high sensitivity and specificity compared to other detection methods, modifications increasing PCR sensitivity may reduce specificity at the same time (cycling conditions, concentration of reagents) except for Nested PCR. Basically, sensitivity is increased by augmented $MgCl_2$ or high primer and polymerase concentrations while specificity is decreased.

Nested PCR technique is widely established for high sensitivity PCR with simultaneously increased specificity. The technique usually includes two primer pairs that are applied in two consecutive cycling reactions. During the first round a relatively large amplicon is produced with the first primer pair. This PCR may be of very low efficiency so that the PCR product can not be detected by electrophoresis. However, even small volumes of the reaction mix contain a great number of DNA molecules after amplification. These PCR products are then added as template to the second PCR. The second primer pair matches to an internal region of the amplicon and larger amounts of a second smaller PCR product are generated.

Nested PCR. Nested PCR technique is on the one hand used for avoiding effects of polymerase inhibitors (Section Inhibitors) but also for enhancing sensitivity. As an additional effect specificity is also increased, since the second primer pair would not match to unspecific amplicons from the first round but solely to specific PCR products with complementary sequences. Beside these positive effects, extreme sensitivity also means increased risk of false positive reactions due to minute contaminations.

Whole genome amplification (WGA). WGA is an in vitro procedure to amplify whole genomic DNA from trace quantities of starting material. This procedure is a tool applied in many fields (genetic analysis, linkage studies, pharmacogenetics, human disease research) and is especially valuable in cases where only limited amounts of precious material are available.

PCR-based and non-PCR-based techniques were developed to amplify whole genomes. PCR-based methods include degenerate oligonucleotide primed-PCR (DOP-PCR) and primer extension PCR (PEP). Non-PCR-based methods are multiple displacement amplification WGA (MDA WGA) and T7-based linear amplification (TLAD).

DOP-PCR uses a partially degenerate oligonucleotide primer that binds during several low-temperature annealing cycles at many sites in the genome (approximately 10^6 sites in humans). This preamplification step is followed by PCR amplification at a higher temperature to allow only those initial fragments that are tagged with the specific part of the primer sequence to be amplified.

PEP uses a 15 base random oligonucleotide primer which anneals at a low stringency temperature (37°C) to its target. Subsequently the annealing temperature is gradually ramped (10 s/°C) up to 55°C prior to a 4 min elongation step (55°C). Using this technique approximately 78% of the genome from a single haploid cell can be copied at least 30 times during 50 cycles of amplification.

MDA WGA is based on the unique property of Phi29 DNA polymerase to replicate linear genomic DNAs by strand displacement amplification. The reaction uses random hexamers as primers for DNA synthesis and is performed routinely overnight at 30°C. This powerful technology allows up to 1000-fold amplification of starting material.

The first step in the TLAD protocol implies digestion of genomic DNA with the restriction endonuclease AluI. Terminal transferase is then used to add polyT tails to the 3'-ends of the digested DNA strands. An oligonucleotide primer, containing a 5' T7 promoter and a 3' polyA tract, is annealed to the genomic fragments, and Klenow is used for synthesis of the second strand. The reaction products of this synthesis are then used as templates for the in vitro transcription reaction.

PCR-based methods are generally more amenable to non-specific amplification artifacts and an incomplete coverage of loci (allelic bias). PCR products have a mean length of about 3 kb (kilobase pairs).

One of the major advantages of non-PCR-based methods is that they do not introduce the sequence- and length-dependent biases observed in PCR-based amplification. Amplification products generated with isothermal methods are about 10 kb in length.

[Top of Page]

9. Contamination

9.1. Sources of Contamination

Since PCR is the most sensitive method so far for DNA/RNA detection even minute contamination of reagents, PCR cups, pipette tips etc., with template DNA may be the reason for false positive PCR results. Amplicons are the most probable source for contamination of PCR reactions. For example, if the product yield after PCR is about 1 µg, a droplet volume of 1 pL (10^{-12} L) would be enough to contaminate a PCR reaction with 1×10^5 amplicon molecules assuming an amplicon size of 500 bp.

Furthermore, samples containing DNA templates to be detected may be a source of contamination. Therefore, PCR cups should be closed as soon as sample DNA is added to the PCR reaction. To be on the safe side, samples as well as positive and negative controls should be amplified in duplicate or even in triplicate.

Depending on the kind and properties of the template particular precautions have to be taken. Samples which are to be examined for human sequences must be considered as especially problematic since template DNA is ubiquitous, e.g., as hair and/or skin cells, as human cells on surfaces or even in the air. Also bacteria that are commensals on the human skin

(*Staphylococcus* spp., *Streptococcus.* spp., *Mycobacteria* spp.) require special measures.

9.2. PCR Controls

To avoid false-negative and/or false-positive results controls have to be included. Faulty PCR reactions which would provide a false-negative result can be avoided by *positive controls*. For this purpose the PCR reaction is deliberately contaminated with template DNA. Thus the PCR result must be positive at least for this control reaction otherwise the whole reaction has to be repeated. If inhibitors in the sample are causative for negative results, there are two possibilities to demonstrate this particular problem: (1) add the DNA to be amplified to the sample or (2) add an internal control to the sample.

In both cases the PCR result must be positive if inhibitors do not interfere. In case (2) the internal control must be chosen in such a way that the amplicon sites of internal control and the actual amplicon are different in size. Internal controls can be added to each sample. Thus three different PCR results are possible:

1. Internal control and actual amplicon positive — conclusion: sample positive,
2. Internal control positive, actual amplicon negative — conclusion: sample negative,
3. Internal control and actual amplicon negative — conclusion: sample contains inhibitors

False positive results can be avoided by including *negative controls*. To test PCR reagents for contaminations with template DNA bidistilled water is added to the PCR reaction instead of template DNA. In this case the PCR result has to be negative.

Detection of pathogens in, e.g., blood requires blood samples as negative controls which do not contain the template DNA. Subsequently for DNA preparation from blood samples the negative control is prepared in exactly the same way as regular samples. After DNA preparation the PCR result of this negative control must be negative otherwise the reaction including the preparation of samples must be repeated. The same approach can be used in case of positive controls.

9.3. Decontamination Methods

9.3.1. *Uracil-N-Glycosylase (UNG)*

For this decontamination method dUTP is substituted for dTTP in the PCR reaction. Since affinity of most of the DNA polymerases is similar for both dUTP and dTTP, dUTP is incorporated into the newly synthesized DNA strand during PCR reaction. Thus prior to PCR amplification contaminations by amplicons from previous PCRs can be removed by addition of UNG (AmpErase) to the PCR reaction and incubation of this mixture at 37 °C. UNG catalyzes hydrolysis of uracil-containing nucleic acids and thus destroys uracil-containing amplicons without affecting original template DNA. After hydrolysis UNG is inactivated automatically at the PCR cycle conditions.

This method can be varied in such a way that the PCR primers are synthesized with dUTP instead of dTTP. Such primers are designated as *amplimers*. This modification has the advantage that PCR reactions can be performed with the natural substrate dTTP. Nevertheless, contaminating amplicons are marked by dUTP and hence destroyed by UNG.

Decontamination of PCR reactions by UNG is characterized by the following properties:

1. Efficient inactivation of more than 10^6 template molecules (dsDNA or ssDNA),
2. Efficiency of PCR is not affected,
3. Post-PCR hybridization is not influenced,
4. No additional equipment is necessary, and
5. The PCR product can be sequenced and cloned regularly

However, the following PCR manipulations of the amplicon are not possible:

1. dUTP does not serve as substrate for endonucleases
2. Only G – C parts of amplicon sequence can be used in restriction analysis
3. Protein – DNA interaction studies may be impaired

9.3.2. *UV Light*

Contaminating DNA in PCR cups and reagent components except sample DNA, primers, and DNA polymerases can be removed by UV irradiation. In this context several properties not only of the DNA but also of the UV light have to be taken into consideration:

1. Pyrimidine base content of the DNA, especially thymidine content,
2. Size of the DNA,
3. Irradiation time,
4. Distance from the UV source,
5. Wavelength of the UV light, and
6. Energy of the UV light

A transilluminator may be used as UV source but to obtain reproducible results, items 3 to 6 need to be controlled. Several UV sources (UV cross-linkers) are available which allow exact control of these items (Amersham, Stratagene). To achieve effective decontamination, PCR cups and solutions should be treated for more than 10 min at 254 nm wavelength and with a power of 100 µJ/min.

9.3.3. *Enzymes*

In cases where other methods are inappropriate, DNA modifying enzymes like DNAse I, exonuclease III and restriction enzymes may be used for decontamination. If for example solutions (e.g., media) are to be used which are not specified as DNA free and which are sensitive to other decontamination methods mentioned above, enzymes can be applied to "sterilize" these solutions. Subsequently, added enzymes are inactivated by heating solutions to 100 °C for 10 min or by autoclaving.

9.3.4. *Psoralens and Isopsoralens*

Prior to PCR contaminating DNA can be "inactivated" by *psoralens* (natural furocoumarins, e.g., 8-methoxypsoralen) which irreversibly cross-links both DNA strands during irradiation with UV light (wavelength 300 – 400 nm). By linking of both strands denaturing of DNA and thus amplification is no longer possible. Psoralens are added to PCR reactions (without sample DNA) and contaminating DNA is "inactivated".

Isopsoralens (4-aminomethyl-4,5-dimethylisopsoralen and 6-aminomethyl-4,5-dimethylisopsoralen) form DNA-monoadducts upon exposure to long-wave UV light. Such modified DNA is no longer recognized as substrate by DNA polymerases. Isopsoralens are added to the PCR reaction prior to amplification but irradiation with UV light is performed after amplification so that amplicons are inactivated at once.

[Top of Page]

10. Quantitative PCR

Soon after the development of PCR experiments were performed not only to qualify but also to quantify nucleic acids in samples. Applications of quantitative methods are numerous and include diagnostic testing as well as therapeutic monitoring for neoplastic or infectious diseases or analysis of gene expression (Section PCR for Analysis of Gene Expression) [34], [35].

Quantitative PCR is used to infer from amplified products to the number of template molecules prior to PCR. The relation between the initial number of molecules and amplified product is described by an exponential function:

$$y = a(1 + E)^n$$

with

y = product yield (amount of DNA after amplification)
a = amount of template DNA prior to PCR
E = efficiency of the amplification process
n = number of cycles

Logarithmic transformation of this equation results in an equation for a straight line:

$$\log\ y = \log\ a + n\ \log(1 + E)$$

with log *a* being the *y*-axis intercept and log (1 + *E*) the slope of the straight line. This equation can only be applied to the exponential phase of the PCR reaction which depends on the initial number of template molecules and the number of PCR cycles.

To quantitate template DNA, a standard control DNA with a known number of molecules has to be coamplified. Control DNA may be coamplified in a separate PCR cup (external control) or in the same PCR cup as the sample DNA (internal control).

10.1. External Control

External controls are amplified by subjecting different amounts of a known DNA to PCR to obtain a standard curve. The amount of sample DNA amplified in parallel can be derived directly by interpolation of the standard curve. Preconditions for this quantification method is that the efficiency of both PCRs, that of the control DNA and that of the sample DNA, is identical and that the amplification reaction is in the exponential phase.

10.2. Internal Control

PCRs using internal controls (IC) for quantification may be performed as multiplex or competitive PCR.

In *multiplex PCR* the primer pair for the IC is derived from a different DNA molecule than the primer pair for sample DNA whereas in *competitive PCR* the same primer pair is used for both, the IC and sample DNA. Preconditions for quantification by multiplex PCR are the same as for PCRs using external controls. The only advantage of multiplex PCR as compared to PCRs using external controls is that both IC DNA and sample DNA are amplified in the same tube and thus problems due to variations in tube consistency or position in the thermal cycler are eliminated.

Competitive PCR. In competitive PCR amplicons generated from the IC and sample DNA by the same primer pair have to be differentiated. This can be achieved by modification of IC DNA either by deletion or insertion of an internal sequence. If an internal region is deleted amplicons generated from the IC are smaller and if an internal region is inserted they are larger than amplicons from sample DNA. In competitive PCR both IC DNA and sample DNA compete for the same primer pair. If the amount of initial template DNAs (IC and sample DNA) is unequal, PCR product yield is shifted to the credit of the DNA being in excess. Only when the amount of initial template DNA is equimolar, product yield for IC and sample DNA is identical. Thus the amount of sample DNA prior to PCR can be directly calculated from the known amount of IC DNA

(Fig. 6). If PCR products have been labeled during amplification (e.g., by incorporation of radiolabeled nucleotides, fluorescent or chemiluminescent labeled primers) gels can be analyzed by densitometric methods. In contrast to the above-mentioned quantification methods, quantification by competitive PCR is independent of cycle number and thus independent from the amplification phase.

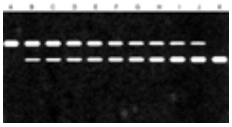

Figure 6. Schematic representation of competitive PCR for quantification of PCR products. Sample (upper bands) and internal control (lower bands) are amplified at different dilutions (lane A – K). Identical band intensities of sample and control DNA indicate identical starting copy numbers of template DNA in the sample and initial copy number of the internal control (lane F)

10.3. Real-Time Quantification

The breakthrough in PCR quantification was achieved by the development of systems making it possible to detect PCR products immediately during their formation in the reaction tubes. The fluorescence signal produced during PCR can be measured on-line in an integrated thermocycler by an automatic system for quantitative PCR. Meanwhile different systems are available for realtime quantitative PCR (Table 3).

Real-time quantification of PCR products is possible with fluorescent dyes binding to dsDNA. For this purpose Sybr Green I, which is highly sensitive and specific for dsDNA, is particularly suited. Its fluorescence is greatly enhanced by binding to dsDNA and the emitted signal is proportional to newly synthesized dsDNA.

10.3.1. *Fluorescence Resonance Energy Transfer (FRET)*

If more specific detection is required, adjacent hybridization probes can monitor the appearance of a specific product during the reaction. The probes are designed to be internal to the amplification primers. The 3′ end of the probe is labeled with a donor fluor while the 5′ end of an adjacent probe is labeled with an acceptor fluor. FRET will occur only when the two probes hybridize to the amplicon in close proximity. During FRET the donor fluorophore is excited by an external light source and emits light that is absorbed by a second acceptor fluorophore (Fig. 7). The acceptor fluorophore emits light of a different wave length that is filtered and measured. The amount of FRET signal is proportional to the amount of specific PCR product available for hybridization, thus the signal increases after each thermal cycle. The FRET principle for fluorescent PCR analysis on the LightCycler is called hybridization probes format.

Figure 7. Schematic representation of the Fluorescence Resonance Energy Transfer (FRET) principle. A donor fluorophore (F1) is excited by an external light source. Emitted light is absorbed by an acceptor fluorophore (F2), that is in close proximity. The acceptor fluorophore then emits light of a different wavelength that can be measured.

The LightCycler instrument (Roche Diagnostics) (Fig. 8) is a rapid thermal cycler combined with a microvolume fluorimeter utilizing high quality optics. This machine features rapid thermal cycling and flexible on-line, real-time detection of the reaction kinetics. Heating and cooling are controlled by alternating heated and ambient air as the medium for temperature transfer. Due to the low mass of this medium, very rapid temperature exchanges can be achieved within the thermal chamber. The LightCycler amplification technique uses durable glass capillaries as reaction vessels, placed in a sample carousel. Because of the high surface-area-to-volume ratio, capillaries are highly efficient at transferring heat, thus permitting rapid cycling. With the LightCycler 30 cycles can be performed in less than 30 min. Furthermore the glass capillary serves as an optical element for signal collection (similar to glass fiber optics), piping the light and concentrating the signal at the tip of the capillary. The effect is efficient illumination and fluorescence monitoring of the microvolume samples. The capillaries permit reaction volumes from 5 – 20 µL.

The fluorimeter, as part of the machine, houses the light source: a maintenance-free light emitting diode (LED). The emitted light from the diode is filtered and focused by an epi-illumination lens onto the capillary tip. Fluorescent light emitted from the sample is then focused by the same lens, passed through a dichroic mirror, filtered appropriately and focused on data collecting photoelectric cells. Data may be collected once per cycle or continuously for quantification of PCR products. For confirmation of specificity, data are collected stepwise after defined temperature intervals so as to monitor melting characteristics of the amplicon.

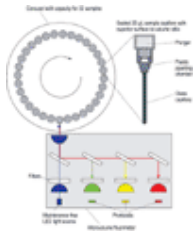

Figure 8. Principle of the LightCycler

10.3.2. *Förster Energy Transfer (TaqMan System)*

The 5′-3′-exonuclease activity of the *Taq* DNA polymerase may also be used to generate signals during PCR. For that purpose a probe labeled with two fluorescent dyes is added to the template DNA. The probe contains a reporter dye at the 5′ end and a quencher molecule at the 3′ end. The vicinity of the quencher molecule reduces the emitted fluorescence of the reporter dye by Förster energy transfer. During PCR the probe anneals to the denatured template DNA and is subsequently hydrolyzed through the 5′-3′ exonuclease activity of the polymerase. Thus reporter and quencher dye are separated from each other and the reporter signal increases (Fig. 9). Synthesis of the complementary strand is finished

subsequently. Since in each cycle reporter molecules are liberated, fluorescence increases proportional to the number of synthesized PCR products. For quantification of results the increase in reporter fluorescence is measured real-time during the whole PCR process. For qualitative aspects, an end point determination is performed. For application of techniques based on Förster′s energy transfer refined technical equipment has been developed.

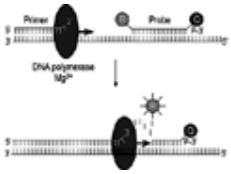

Figure 9. Förster Energy Transfer (FET) principle. PCR containing two primers and one probe is performed. The probe is labeled by a reporter (R) and a quencher (Q) molecule and is phosphorylated at its 3′ end. Phosphorylation avoids strand elongation by the DNA polymerase. As long as reporter and quencher are in close proximity fluorescence of the reporter dye is quenched. As soon as the reporter-labeled nucleotide is removed by the 5′-3′ exonuclease activity of *Taq* DNA polymerase the quenching effect is neutralized and the fluorescence of the reporter dye can be emitted.

In the TaqMan system (Applied Biosystems) blue laser light is lead through 96 glass fibers to the sample position of the integrated thermal cycler for fluorescence excitation (Fig. 10). Emitted fluorescence is directed through the glass fibers over a spectrograph to a charge-coupled device (CCD) camera recording every few seconds an emission spectrum of each probe in the microtiter plate.

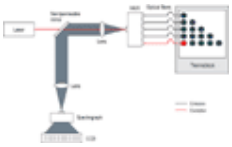

Figure 10. Principle of the TaqMan system

MUX = Multiplexer; CCD = charge-coupled device

Recording of spectra makes it possible to use a second TaqMan probe labeled with a different reporter dye for the qualitative or quantitative measurement of an internal control. The possibility to directly follow up the PCR process allows a novel method for quantification of nucleic acids. Different PCR reactions are characterized by the time to their first detection and not as usually by the amount of accumulated PCR product. The more starting material is subjected to PCR the earlier a significant increase in fluorescence is measured. The C_T value (threshold cycle) characterized by the number of cycles at which the fluorescent signal exceeds a fixed threshold serves as parameter to determine the number of template molecules prior to PCR. For different PCRs, standard curves have to be generated by plotting C_T values versus initial copy numbers (Fig. 11). Quantification of the copy number of an unknown sample is achieved by extrapolation based on such standard curves.

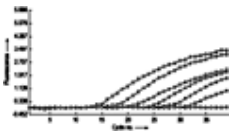

Figure 11. Typical curves obtained from the TaqMan system to detect C_T (threshold) values.

Since the whole reaction is monitored, the linear range of the C_T curve (C_T values versus initial copy numbers) is more than six orders of magnitude. Post-PCR handling of samples is omitted; thus sample throughput can be increased and the contamination risk is decreased dramatically. Automation of the quantification process simplifies many applications as there are quantification of pathogens or transcription analysis in fields like medical research or pharmaceutical development.

Detection of PCR products in closed reaction tubes without further manipulation of amplicons is not only advantageous in quantitative PCR but also in end point analysis (5′ nuclease assay) to detect pathogens in food or drinking water.

[Top of Page]

11. Applications

11.1. Detection of Pathogenic Agents

Diagnostic PCRs offer several advantages over classical diagnostic methods: in vitro cultivation and differentiation of pathogens are hazardous, time consuming, or impossible [36]. Additionally, cultivation is in some cases restricted to certain laboratories. Furthermore, samples may be too small or contaminated for cultivation and sometimes alternative serological methods do not exist. Altogether the use of nucleic acid amplification techniques within the medical microbiology laboratory is becoming more and more accepted [37-39] and there is an increasing number of commercially available PCR test kits (Table 4).

Table 4. Commercially available PCR test kits

Organisms	Test kits	Supplier
Campylobacter jejuni, Escherichia coli O157:H7, Listeria monocytogenes, Salmonella spp. enterica	TaqMan pathogen detection kits	Applied Biosystems

Human, animal, and plant identification	AmpFLSTR amplification kits, StockMarks animal genotyping kit, and TaqMan GMO detection and quantitation System	Applied Biosystems
Viruses: HIV-1, HBV, CMV, HPV	AMPLICOR and COBAS AMPLICOR tests	Roche Diagnostics
Bacteria: *Chlamydia trachomatis, Neissereia gonorrhoeae, Mycobacterium tuberculosis, Mycobacterium intracellulare, Mycobacterium avium*	AMPLICOR and COBAS AMPLICOR tests	Roche Diagnostics
Viruses, bacteria, fungi	Lightcycler reaseach kits	Roche Diagnostics
Campylobacter spp., E. coli O157:H7, Listeria monocytogenes, Salmonella spp.	Probelia	Bio-Rad
Actinobacillus pleuropneumoniae, Borrelia burgdorferi s.l., Campylobacter spp., Chlamydia spp., Clostridium perfringens, Mycobacterium paratuberculosis, Mycoplasma gallisepticum, Mycoplasma synoviae, Pasteurella multocida, Salmonella spp., Yersinia enterocolitica	Bactotype PCR amplification kits	Biotype AG

11.1.1. *Preconditions*

The most important precondition for PCR based detection of microbes is the knowledge of sufficient DNA or RNA sequence data for development of primers (Section Primer). PCR is not convenient for general screening due to its high specificity except for universal PCRs (Section Universal PCR) or multiplex PCR (Section Multiplex PCR).

Generally, PCR based detection of pathogenic agents relies on a two step procedure and applied techniques should be combined in a modular way:

1. Isolation of microbial DNA as template, and
2. Amplification of template DNA by PCR

11.1.2. *Extraction of Template DNA*

Methods for isolation of template DNA are numerous and may be divided into two groups: isolation of total DNA from samples and specific preparation of microbial DNA. Total DNA contains DNA of pathogens plus large amounts of host DNA, that may interfere during PCR. On the other hand total DNA may serve as template for different PCRs detecting various pathogens.

11.1.2.1. *Isolation of Total DNA*

Total DNA from clinical samples is extracted enzymatically or mechanically followed by purification and concentration steps. Preparation methods depend on the sample type and the composition of the expected pathogens' cell wall, respectively. For the detection of intracellular pathogens both, host and microbial cells, must be lysed for DNA extraction. Usually mechanical homogenization (grinding or ultrasound treatment) or differential centrifugation is combined with enzymatic incubation (proteinase K, lysozyme), osmotic cell lysis, or boiling for the extraction of DNA.

After cell lysis added enzymes must be inactivated by boiling or removed by phenol – chloroform extraction. In many cases liberated DNA can be directly subjected to PCR. Nevertheless, if inhibitors (Section Inhibitors) are to be expected or the amount of template DNA is too low DNA must be purified and separated from cell debris.

A widely applied procedure consists in selective adsorption of DNA to silica or glass particles in the presence of high concentrations of chaotropic agents (i.e., sodium iodide, sodium perchlorate, or guanidinium thiocyanate) [40-42]. Elution of DNA is achieved with low-salt solutions or water. Selective adsorption of DNA to silica matrices allows removal of PCR inhibitors (Section Inhibitors). For purification of large amounts of DNA diatomaceous cell walls (petrified cell walls of algae almost entirely consisting of silicon dioxide and larger than synthetic silica particles) are suitable, because insoluble complexes of DNA and silica are avoided [43]. However, commercial silica matrices and convenient buffer systems, including lysis buffer, are widely distributed (e.g., BioRad, Qiagen, Eppendorf).

11.1.2.2. *Isolation of Pathogen Specific DNA*

For this purpose three techniques are established:

1. Propagation of pathogen (mainly on agar plates) and lysis of single colonies,
2. Immunomagnetic separation (IMS), and
3. Separation of template DNA by paramagnetic beads coated with a template specific DNA probe

The first procedure is mainly used for pathogens that are easy to cultivate. In this case PCR is predominantly a tool for further classification and characterization (e.g., slowly growing fungi, toxigenic microbes) and need not be very sensitive.

Immunomagnetic separation (IMS) is based on paramagnetic beads coated with monoclonal antibodies to capture the pathogenic agent specifically. After incubation of samples with coated beads pathogen bearing beads may be separated

magnetically and hence isolated pathogens are lysed for extraction [44], [45].

Another method, similar to IMS, uses paramagnetic beads coated with oligonucleotide probes complementary to the expected template DNA sequence. Incubation of beads is carried out after lysis of sample materials, pathogen specific DNA is captured by the immobilized DNA probe and magnetically concentrated. Beads may be subjected to PCR without interfering amplification.

The predominant advantages of pathogen specific isolation of DNA are low copurification of PCR inhibitors and generally concentrating of template DNA. All three procedures are not adequate for detection of different pathogens or multiplex PCR.

11.1.3. *Amplification*

Purified DNA may then be subjected to PCR reaction mix. If necessary concentration of DNA is achieved by precipitation with ethanol or isopropanol. For amplification of bacterial, fungal, and protozoal DNA conventional PCR in tubes is normally applied. Even samples with DNA viruses may be treated accordingly. Only amplification of RNA virus sequences needs reverse transcription of RNA to cDNA prior to PCR, using a commercially available reverse transcriptase.

Just as polymerases reverse transcriptases need starter molecules to prime the reaction. Usually nucleotide random hexamers or primers are used and the reaction mix is incubated at 42 °C for 45 min. Reverse transcription followed by PCR is often designated RT-PCR, and several attempts have been made to combine both reverse transcription and amplification in one tube.

Tth polymerase, isolated from *Thermophilus thermophilus*, has been shown to exhibit reverse transcriptase activity in the presence of manganese ions and polymerase activity with Mg^{2+} ions as cofactors. This enzyme has become available as recombinant protein r*Tth* and as component of a commercially available RT-PCR kit (Applied Biosystems). In contrast to other reverse transcriptases with an optimal activity at 42 °C r*Tth* polymerase can be incubated at higher temperatures. This property is particularly important if RNA with marked secondary structure is to be transcribed.

Another possibility for one-step RT-PCR consists in applying viral reverse transcriptase and a Hot-Start DNA polymerase for amplification, since this polymerase may be temperature-activated continuously after reverse transcription. One-step RT-PCR kits are available from many suppliers, e.g. Applied Biosystems, Eppendorf, Invitrogen, Promega, Qiagen, and Stratagene.

11.1.4. *Multiplex PCR*

To reduce expense of time and money, different PCRs may be combined to one reaction designated multiplex PCR [46], [47]. Most important for developing multiplex PCRs is that incorporated primers have very similar physicochemical parameters. Furthermore, the size of amplicons should also be similar, otherwise small amplicons are produced in expense of larger ones. However, a commercial multiplex PCR kit (Qiagen) is available, which contains a unique PCR buffer thus making it possible to perform multiplex PCR without further optimization. Multiplex PCRs are mainly applied in the differentiation of species or subspecies (e.g., characterization of slowly growing fungal isolates in one step) for screening purposes (e.g., pathogens or genetic disorders causing the same clinical signs) [48], and for microsatellite analysis.

11.1.5. *In Situ PCR*

Since the earliest descriptions of PCR technologies, researchers proposed performing amplification reactions in situ in non-disrupted cells and tissues with subsequent detection of the amplicon at the site of origin. The technique of in situ PCR provides a tool to localize rare sequences of DNA or RNA within a heterogeneous population. In situ PCR is widely applied for the detection of pathogens or for histological purposes like tumor diagnostics or genetic disorders [49].

Different types of PCR amplification are possible and the general process is shown in Figure 12.

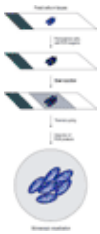

Figure 12. In-situ PCR. Cells or tissues are fixed on slides and lysed. PCR reagents are added and the reaction volume is sealed to avoid evaporation. After thermal cycling PCR products are visualized microscopically.

Basically, direct and indirect in situ PCR technologies exist. For direct in situ PCR incorporation of labels into the amplicon is characteristic, whereas in indirect methods the PCR product is not labeled. Such amplicons must be detected afterwards by in situ hybridization with a labeled probe. For direct labeling of amplicons either hapten-labeled nucleotides (biotin-11-dUTP, biotin-14-dATP, digoxigenin-11-dUTP, fluorescein-15-dATP, numbers are indicating the size of the spacer i.e., the number of CH_2 groups between hapten and nucleotide) or hapten-labeled primers are added to the reaction (⟶ Nucleic Acids – Synthesis of Modified Oligonucleotides). In general, direct methods are faster but less specific. Indirect procedures including additional hybridization steps reduce the problem of unspecific, visible reaction products and are therefore considered the method of choice by many researchers. In situ PCR may also be conducted with RNA as template including reverse transcription step.

Cell smears, cryostat or paraffin sections, whole mounts, cell suspensions or chromosome spreads are considered to be

suitable preparations for in situ amplification. Nevertheless, convenient preparation of samples is crucial for successful in situ PCR protocols. The main difficulty in preparation is that harsh treatment may destroy template DNA or RNA. All preparations are made on silane-coated glass slides. Samples must first be fixed and then cells must be permeabilized. Fixation slides may be immersed in 4 % paraformaldehyde, formalin, ethanol – acetic acid or other fixatives. Formalin-fixed and paraffin-embedded sections (e.g., archival material) must be carefully dewaxed before further preparation. The following pretreatment is necessary in order to achieve penetration of the PCR components into the cells and allow the target sequence to be accessed for amplification. Permeability of cells is achieved by incubation of slides in HCl – pepsin, trypsin, or proteinase K solutions. At the end of incubation, enzymes are destroyed by short boiling. Sometimes additional steps (DNase digestion, removal of endogenous enzymes, etc.) are required prior to PCR. Then the PCR reaction mix is added to the slides which are to be sealed by tight cover discs and clips (AmpliCover discs and AmpliCover clips by PE Biosystems). After assembly the slides are inserted into particular thermocycler blocks and cycling is performed with two or three temperature programs. After amplification, the tissue sections can be refixed (ethanol, paraformaldehyde) to prevent loss of amplified DNA. Finally direct microscopic investigation or hybridization are performed.

11.1.6. *Universal PCR*

Several attempts have been made to develop so-called universal PCRs, so as to detect every kind of pathogen by a singular PCR system. All these experimental systems are based on amplification of highly conserved genetic regions common to all pathogens. Amplification is usually combined with sequencing of PCR products in order to find characteristic differences between species. Widely used genes for this purpose are 16S rRNA genes [50-52] or spacer regions for the detection and differentiation of bacteria [50], [53], [54]. Other sets of universal primers are based on genes of cold-shock proteins [55] or ribosomal proteins [56].

11.1.7. *PCR Controls for Diagnostic Purposes*

Because of the importance and implications of diagnostic results for therapy, prognoses, or epidemiology each reaction step must be controlled thoroughly. False positive or false negative results are mainly caused by contamination and inhibitors, respectively. To avoid these, the control reactions described in Section PCR Controls should be carried out.

Although PCR is highly specific post-PCR analysis is required in diagnostics to confirm the correctness of amplicons. Restriction analysis (Section Restriction Analysis), hybridization with specific DNA probes (Section Hybridization) and sequencing of the amplicon (Section Sequencing) prove specificity. Hybridization for this purpose may be carried out as microplate-capture method, which increases the sensitivity of the procedure and allows semi-quantitative evaluation [57], [58]. Altogether, diagnostic PCR procedures including template preparation should not take more than 24 h.

11.2. Detection of Genetic Disorders or Cancerous Cells

PCR technologies allow precise identification and analysis of mutations resulting in genetic disorders or neoplastic diseases. Reliable diagnosis of such mutations is particularly important for persons suffering from genetic diseases and individuals carrying pathogenic alleles, for prenatal consulting or tumor diagnostics. Special attention must be paid to the reliability of PCR results. Furthermore, results must be interpreted in the correct clinical context [59]. To save time in diagnosis multiplex PCRs (Section Multiplex PCR) may be developed.

Due to its high sensitivity and specificity even small amounts of altered DNA can be detected by PCR which is of particular interest in tumor diagnostics. Especially ARMS procedures (Section Known Mutations) are suitable since mutated DNA can be amplified out of a high background of intact DNA. Generally, methods for detection of known and unknown disorders must be distinguished.

11.2.1. *Known Mutations*

11.2.1.1. *Amplification Refractory Mutation System (ARMS)*

In this procedure, DNA is amplified with two primer pairs in separate assays. The first primer pair contains one primer specific for the mutated allele: the mutation is part of the primer sequence. With this primer pair only mutated DNA is amplified. The second primer pair amplifies the unaffected allele. Examination of diploid DNA leads to four possible results:

1. Amplification only with primer pair 1: DNA only contains mutated alleles. The patient is homozygous for that mutation.
2. Amplification only with primer pair 2: DNA only contains unaffected allele. The patient is homozygous.
3. Amplification with both primer pairs: DNA contains both alleles. Patient is heterozygous.
4. Amplification only with primer pair 2, but symptoms in patient: another mutation is responsible for disease.

In the case of cancer patients the results look like those of above-mentioned heterozygous persons. Combined techniques based on quantitative PCR and ARMS have been developed [60]. ARMS has been applied for example in the prenatal diagnostics of hereditary diseases and in screening for cancer cells during or after cancer therapy.

11.2.1.2. *Allele-Specific Oligonucleotides (ASOs)*

Similar to confirmation procedures for PCR-based detection of microbes, in this method PCR is followed by hybridization. Allele-specific oligonucleotides are short, single-stranded oligonucleotide probes. Differences in nucleotide sequence may be as small as one base pair so that point mutations may be detected. ASO-tests are usually conducted as dot blots (⟶ Nucleic Acids – Hybridization Techniques for Nucleic Acid Detection) with allele-specific oligonucleotides as probes.

If several mutations are to be examined at once, different allele-specific oligonucleotide probes can be immobilized to membranes. Then hybridization is carried out as reverse dot blot with labeled amplicon.

11.2.1.3. *Introduction of Restriction Sites*

Mutations may also be detected by enzymatic restriction (Section Restriction Analysis) of amplicons. Either mutated DNA

together with a primer constitutes a new restriction site or the pathogen allele contains a modified restriction site. Amplified DNA of pathogen alleles is then restricted in a different manner compared to DNA of normal alleles which easily may be shown in electrophoresis. However, partial restriction of DNA may simulate heterozygous genotype of patients.

11.2.1.4. *Competitive PCR*

For this technique allele-specific labeled primers are chosen. During amplification amplicons of identical length for every allele are generated that may be distinguished by different primer labeling (i.e., fluorescence). Amplicons of homozygous patients show the original fluorescence of primers whereas amplicons of heterozygous individuals show mixed fluorescence.

11.2.1.5. *Primer Extension PCR (PEST)*

This method represents a modified ARMS procedure. To one primer pair complementary to the interesting DNA region a third primer is added that corresponds to mutated DNA. In homozygous individuals with unmodified alleles only one large amplicon is generated. In heterozygous or homozygous individuals with mutated DNA two different amplicons are generated (Fig. 13).

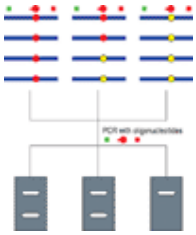

Figure 13. PEST PCR

A) DNA molecules consisting of genes of identical allelic type (indicated by red circles) are subjected to PCR with primers specific for the mutated (red circle) gene and two primers from conserved regions of the gene (blue and green rectangles). Two PCR products are generated, one from the two conserved primers (larger amplicon) and one from the mutation-specific primer and the conserved primer (smaller amplicon).

B) In the case of heterozygous DNA the same PCR result is obtained as in panel A.

C) With DNA containing the unmutated gene (yellow circles) only the larger PCR product with the two conserved primers is obtained. The mutation-specific primer does not match.

11.2.2. ***Unknown Mutations***

For analyzing unknown mutations electrophoretic characteristics of differing PCR products are used.

11.2.2.1. *Denaturing Gradient Gel Electrophoresis (DGGE)*

This method allows the separation of DNA molecules, i.e., PCR products, differing in as little as a single base pair. The differentiation is based on the melting properties or denaturation characteristics of DNA. Amplicons are therefore analyzed on gel matrices with increasing denaturant concentration (formamide, urea, temperature). Parallel and perpendicular DGGE are distinguished depending on the orientation of the denaturant gradient.

In non-denaturing gel electrophoresis amplicons derived from different alleles migrate as one band whereas in parallel DGGE two bands are visible. Perpendicular DGGE however results in characteristic sigmoidal curves (Fig. 14).

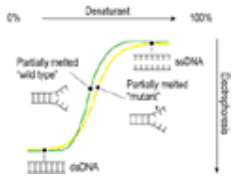

Figure 14. Perpendicular DGGE. Amplicons from mutant (green line) and wild type DNA (yellow line) are analyzed on a gel with horizontally increasing denaturant concentration (0 – 100 %). Electrophoresis is performed perpendicular to the gradient. At low denaturant concentrations dsDNA is maintained. At medium concentrations of denaturant both DNAs, mutant and wild type DNA are partially melted whereas wild type DNA melts at slightly lower denaturant concentrations than mutant DNA. At denaturant concentrations near 100 % mutant and wild type DNA are completely denatured to ssDNA.

11.2.2.2. *Single Strand Conformation Polymorphism (SSCP)*

Differences in single-stranded DNA of up to 400 bp in length may be detected during electrophoresis. For comparison of putative DNA alterations PCR products are restricted, denatured, and separated electrophoretically as single-stranded DNA.

11.2.2.3. *Chemical Cleavage of Mismatches (CCM)*

Amplified DNA from altered alleles and labeled amplicons of normal alleles are mixed and denatured. During denaturation so-called heteroduplex DNA is generated that contains one marked strand derived from normal DNA and one altered strand. In the mutated region mismatching occurs. Mismatched DNA is cleaved chemically (piperidine potassium permanganate), eventually after labeling of mismatched bases by osmium tetroxide or hydroxylamine. Cleaved PCR products are shown afterwards by electrophoresis in the case of DNA alterations.

Although the described methods have been well established for analyzing genetic disorders or cancerous mutations, automatic sequencing (Section Sequencing) as alternative procedure is becoming more and more important and is considered to be the safest way to detect mutations [61].

11.3. PCR Analysis in Evolution and Taxonomy

Apart from characterizing DNA of individuals (see also Section Detection of Genetic Disorders or Cancerous Cells), PCR technology may also be applied for epidemiological purposes and phylogenetic or phylogeographic studies [62]. Generally, similarity or diversity of organisms is to be shown and measured.

However, precondition for analyzing genetic relationships is the occurrence of polymorphisms. Polymorphisms are characteristics (i.e., DNA sequences) that evolve continuously or exist in different forms. Examples for commonly known polymorphisms are blood group antigens, isoenzymes, human leukocyte antigen (HLA) antigens, satellite DNA [mini- and

microsatellites, variable number of tandem repeats (VNTRs)].

Depending on the period that is to be examined (few generations, centuries or thousands of years) different kinds of polymorphisms must be examined. For short periods or high resolving procedures characteristics that develop fast are studied, e.g., DNA sequence of introns, mitochondrial DNA [63], [64] or microsatellites. For examinations spanning large periods DNA sequences that evolve slowly are analyzed, for example, extremely conserved t- or rRNA genes. Particularly PCR-based analysis of rRNA genes or ribosomal spacer regions have become an important tool for taxonomic questions and differentiation of species or subspecies (see also Section Universal PCR) [53], [54], [65].

Commonly applied methods for studying relationship are blood group serology, analysis of amino acid sequences, restriction fragment length polymorphism (RFLP) [66], PCR typing of microsatellite or HLA genes, random amplified polymorphic DNA (RAPD), or DNA sequencing. RFLP, PCR typing, and RAPD will be discussed in the following sections and may be summarized as fingerprint methods, since they all lead to a “bar code”-like result, the genetic fingerprint [67]. Altogether, fingerprinting methods are an extremely useful tool for differentiation and characterization of organisms.

11.3.1. *Restriction Fragment Length Polymorphism*

For RFLP DNA (chromosomal, plasmid DNA, or PCR amplified DNA) is restricted with endonucleases, so that different DNA fragments are generated that can be separated electrophoretically. Fragment polymorphism is then visualized by staining DNA with ethidium bromide or by hybridization with labeled probes. Banding patterns are supposed to show characteristic differences [51], [68].

A very common application of RFLP fingerprints is analysis of eucaryotic DNA microsatellites. Satellite DNA is therefore amplified and marked as probe for hybridization with restricted and electrophoretically separated genomic DNA. Resulting hybridization patterns are characteristic for each individual.

11.3.2. *PCR for Typing Satellites or HLA Genes*

Satellites consist of repetitive DNA sequences (simple mononucleotide to pentanucleotide repeats) of variable number and have become the marker of choice in many research fields in biology, animal and plant breeding, such as genome mapping, population studies, testing of parentage and kindship, cancer diagnostics as well as forensic identification [69]. Microsatellites are abundant, codominant, highly polymorphic and spread across the whole genome.

One of the major fields of application is linkage analysis, where microsatellite markers, whose alleles are seen preferentially by those members of families with the trait of interest (e.g., milk yield, disease, disease resistance) are said to be linked and should therefore colocalize with the responsible gene.

On the other hand, the alleles of unlinked markers should be inherited equally between the different phenotypes. The likelihood that any preferential inheritance seen for a certain marker is the result of genuine linkage as opposed to simple chance is expressed as LOD scores. Requirements for linkage analysis increase with the number of loci being involved in the disorder or phenotypic trait.

Complex traits that are not inherited in a simple Mendelian fashion but by several to many loci with smaller effects on the trait exhibit a great challenge on family numbers and statistical analysis involved. Such traits are called “quantitative trait loci” (QTL). They can be studied by QTL analysis [70].

Usually many loci (microsatellites) are analyzed in parallel and thus multiplex PCR assays are a must for such type of analysis. Multiplex PCR products generated not only differ in length, but also in the fluorophores covalently attached to the 5'-end of forward or reverse primers. Thus even PCR products having the same length may be differentiated by their fluorophores. Amplification products are separated by high-resolution polyacrylamide gel electrophoresis, allowing one single nucleotide length difference to be analyzed. Laser-induced fluorescence signals of PCR products allows for online monitoring of electrophoresis patterns. Amplification generates characteristic sets of PCR products for individuals or groups (Fig. 15). Typing of HLA genes is described in Section Tissue Typing.

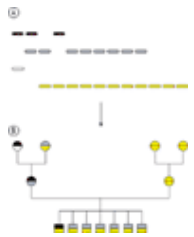

Figure 15. Typing of microsatellites from three generations of one family by fingerprint analysis and deviated family tree. Maternal grandmother (lane 1) and grandfather (lane 3) possess two different alleles (black and white; grey and yellow). Paternal grandmother (lane 11) and grandfather (lane 13) are homozygous (yellow). Consequently the father (lane 12) is also homozygous (yellow). The mother (lane 2) is heterozygous (black and grey). Progenies are heterozygous possessing all but one (black and yellow) identical alleles (grey and yellow).

11.3.3. *Random Amplified Polymorphic DNA*

RAPD PCR is based on unspecific binding of primers to unknown DNA sequences. Primers for this purpose are short and composed randomly. Unspecific amplicons generated initially during RAPD reactions are later on amplified specifically, so that a characteristic set of distinct PCR products is generated (Fig. 16). A typical temperature profile for RAPD consists of two initial cycles with extremely low annealing temperatures (40 °C) for several minutes followed by 20 – 30 regular PCR cycles. Furthermore, Mg^{2+} ion concentration may be elevated [71], [72]. RAPD techniques are widely used in plant breeding research and for characterization of bacterial isolates [73]. One of the major problems of this technique is low reproducibility [74] of results and low discriminatory power [75], [76]. RAPD techniques are also designated as arbitrarily primed PCR (AP-PCR).

PCR fingerprinting methods similar to RAPD are ERIC-PCR (enterobacterial repetitive intergenic consensus PCR) [77], DAF (DNA amplification fingerprinting) [78] and REP-PCR (repetitive extragenic palindromic PCR) [79], [80].

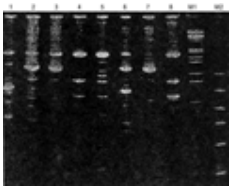

Figure 16. PCR fingerprinting analysis. Different species (lane 1 – 8) have been subjected to DAF PCR showing different patterns on an agarose gel. M1, M2 molecular weight marker.

11.3.4. *Analysis of Fingerprints*

Besides visual analysis, many computer-based image analyzing systems exist. Comparison of fingerprint patterns and quantifying of similarity as a measure for relatedness are based on mathematical methods ["Dice coefficient" [81], UPGMA (unweighted pair group method, averages)] [82]. Computer software packages like PHYLIP (Felsenstein) simplify analysis and help to construct figures of relationships (dendrograms). A precondition to evaluate time course in phylogeny is the knowledge of mutational rates of the examined loci.

11.3.5. *Restrictions and Future Aspects*

Since the beginning of the 1990s developments in automatic DNA sequencing reduced the importance of described fingerprinting methods. Comparison of genetics may also be performed on the basis of completely sequenced genes [83]. The only precondition for routine and comparative sequencing is that sufficiently conserved DNA regions exist for the construction of primers. Sequencing of PCR products for analyzing relationship altogether allows detection and characterization of smallest polymorphism, i.e., point mutations.

PCR based fingerprinting methods became after all an important tool for molecular archeology. To reduce DNA loss due to aging, particularly short DNA fragments are amplified that existed originally in several copies.

Two well known examples for PCR based phylogenetic studies are rRNA gene analysis for taxonomic purposes or high resolving mt (mitochondrial) DNA analysis for identification of singular (vertebrate) species.

11.4. Tissue Typing

Major histocompatibility complex (MHC) molecules, belonging to three different groups of cell surface antigens, are coded by extremely polymorphic genes (10 – 50 alleles per locus) in the HLA (human leukocyte antigen) region of human chromosome 6. Due to high genetic variability recombination results in enormous polymorphisms so that every individual may be identified by its individual MHC antigen combination.

MHC I molecules are expressed on the surface of virtually every cell type containing nuclei. They represent the classical transplantation antigens and may play a key role in the activation of killer cells. MHC II molecules are mainly expressed by cells belonging to the immune system. These cells internalize antigens and present them to others for activation of immunological reactions.

One of the major problems in transplantation medicine is identification of organs as "not-self" by the immune system followed by destruction. To avoid this reaction high similarity in MHC I antigens and HLA loci is needed. Typing of HLA loci is also applied for forensic purposes to identify offenders. Meanwhile convenient commercial kits for tissue typing have become available.

Furthermore, several diseases are described to be associated with certain HLA types, so that identification of HLA types may indicate an elevated risk to suffer from these diseases and accelerate diagnosis.

Besides serological methods for differentiation of MHC antigens, PCR with consecutive analysis of amplicons is becoming more important [84]. By this way antigen differences are not only detected but also characterized immediately.

After amplification of HLA genes, amplicons may be characterized by three different procedures:

1. Enzymatic restriction (Section Restriction Analysis),
2. Denaturing gradient gel electrophoresis (DGGE) (Section Denaturing Gradient Gel Electrophoresis (DGGE)), or
3. Hybridization with allele-specific oligonucleotides (ASO) (Section Allele-Specific Oligonucleotides (ASOs)) [42]

11.5. PCR in Food Analysis

During public discussions concerning production and trade of genetically modified food, the necessity of sensitive and specific detection methods for modified food became evident. With these methods detection of even small amounts of modified food in mixed products must be possible. Today, PCR is well established in food analysis as a most sensitive procedure for detection of genetically modified crops [85] or microbial contamination (Section Detection of Pathogenic Agents).

The detection of genetically modified organisms (GMOs) is only possible if the genetic modification, i.e., sequence is known. For the detection of GMOs in the food sector the types of sequences listed in Table 5 have been applied if intact recombinant DNA is available.

Table 5. Examples for sequences which may be used to detect genetically engineered food

Selective marker genes	
npt II:	neomycin-/kanamycin-/ G 418 resistence (*E. coli*)
hph; hpt:	hygromycin resistence (*E. coli*)
aac C1:	gentamycin resistence (*Micromonospora purpurea*)
cat:	chloramphenicol resistence (*E. coli*)
aad A:	spectinomycin resistence (*E. coli*)
dhfr-Mtx:	dihydroxyfolate (*E. coli,* mouse)
bla:	ampicillin resistence (*E. coli*)
Regulators	
35 S:	35 S promoter, terminator (*CaMV*)
nos:	nopalin synthase gene (promotor and terminator) (*Agrobacterium tumefaciens*)

Official methods now exist for the detection of DNA of glyphosate-tolerant soybeans (Roundup Ready, Monsanto, USA). Glyphosate is a herbicide inhibiting the most important enzyme in the metabolism of aromatic amino acids in plants: Enolpyruvylshikimi-3-phosphate synthase (EPSPS). Glyphosate tolerance of soybeans is caused by a genetic construct coding for EPSPS of *Agrobacterium sp*. and a transfer peptide from *Petunia Hybrida* (transit signal sequence) for introducing of EPSPS into chloroplasts. To detect this genetic modification primer B1 5′-TGA TGT GAT ATC TCC ACT GAC G-3′ and B2 5′-TGT ATC CCT TGA GCC ATG TTG T-3′, derived from the CaMV/35S promoter sequence and the transit signal sequence, respectively, are applied. The recommended annealing temperature for this PCR is 62 °C. PCR results are confirmed by hybridization with a fluorescein-labeled probe (5′-fluorescein-GGG TCT TGC GAA GGA TAG TG-3′).

A major drawback of conventional PCR is lack of accurate quantitative information about the relative GMO concentrations in food mixtures.

Therefore, real-time PCR assays are applied using multiplex PCR techniques [86]. Apart from detection of the genetically modified sequence an endogenous reference gene is detected, allowing for the relative quantification of the GMO content. In this way GMO contents of less than 0.01% could be detected. However, most laboratories now adjust their estimates by comparison to a standard curve, i.e., a series of measurements from a DNA series with known GMO content.

11.6. PCR to Engineer DNA

Techniques such as cloning of genes, modifying genes (mutagenesis), or screening for recombinant clones has been considerably simplified by applying PCR methods.

11.6.1. *Cloning of PCR Products*

Genes of known sequence are easily amplified by PCR out of crude material avoiding the tedious and time-consuming task of screening DNA libraries with gene-specific probes.

Primers used to amplify genes can be designed with additional sequences (e.g., restriction sites) to facilitate insertion of the amplified DNA into cloning vectors. By using two different restriction sites (one in each primer) amplicons can be digested with two different restriction enzymes and thus can be inserted in a directed way into cloning vectors. One should keep in mind that most restriction enzymes do not cut DNA if the recognition sequence constitutes a terminal DNA sequence. Therefore, at least a few more bases should be added 5′ to the actual recognition sequence.

Most DNA polymerases do not generate blunt-ended amplicons but add a 3′-A. These amplicons can be directly inserted into special, T-A vectors engineered for this purpose containing a 5′ protruding T.

Direct cloning of PCR products into blunt-ended cloning vectors is also possible when using DNA polymerases (*Pfu*, Vent) producing blunt-ended amplicons.

Clones obtained by each of these methods may be screened for recombinant DNA using vector specific or insert specific primers.

11.6.2. *PCR Mutagenesis*

PCR is easily applied to generate mutated DNA, i.e., to insert, delete, or substitute DNA sequences or single bases. Mutagenesis is used for example to analyze the effect of mutated regions or even single bases in a gene on the corresponding protein or to investigate structure/function relationships of proteins.

Several methods are available to introduce mutations into a gene [87]. The most widely used PCR technique is overlap extension PCR. It is used to introduce point mutations (Fig. 17), deletions (Fig. 18) or insertions (Fig. 19) into a gene. Four primers (A, B, C, D) are necessary for this approach. Primer sequences C and D are complementary to the same DNA segment but to the opposite strands. Both primers contain a single base mismatched to the target sequence leading to the same sequence alteration (point mutation) in the two primary PCR products. PCR 1 (primers A and C) and PCR 2 (primers B and C) are performed in separate tubes. After amplification excess primers are removed, PCR products from the two PCRs are mixed, denatured and allowed to reanneal. Products which recombine by overlapping of complementary segments are elongated to maintain dsDNA and subsequently these products are amplified with primers A and B. Products from this PCR contain the site-specific mutation. Deletion of DNA segments by overlap extension PCR are

performed with primers C and D complementary to different segments of the target DNA and containing sequences at their 5′ end which are complementary to each other. The deleted DNA segment is flanked by the 5′ end of each primer C and D that matches the target sequences. Two different PCRs are performed in separate tubes, excess primers are removed from the PCR products, mixed and subjected to a PCR with primers A and B. Insertion of small DNA segments by overlap extension PCR are carried out with primers C and D derived from the same target sequence but complementary to opposite strands. At their 5′ end sequences are added which are complementary to each other. The length of these sequences is limited by the oligonucleotide synthesis procedure. After PCR with primers A/C and B/D, performed in separate tubes, both reactions are mixed, excess primers are removed and a PCR with primers A and B is performed. Overlap extension PCR is also used to combine gene fragments from unrelated sequences. In this case primers C and D are derived from different target DNAs and contain additional sequences at the 5′ ends which are complementary to each other. Further manipulations are the same as for introducing insertions (Fig. 19).

Figure 17. Site-specific mutagenesis by overlap extension PCR. Four primers (A, B, C, D) are necessary for this approach. Primer sequences C and D are complementary to the same DNA segment but to the opposite strands. Both primers contain a single base mismatched to the target sequence leading to the same sequence alteration (point mutation) in the two primary PCR products. PCR 1 and PCR 2 are performed in separate tubes. After amplification excess primers are removed, PCR products from the two PCRs are mixed, denatured and allowed to reanneal. Products which recombine by overlapping of complementary segments are elongated to maintain dsDNA and subsequently these products are amplified with primers A and B. Products from this PCR contain the site-specific mutation.

Figure 18. Deletion of DNA segments by overlap extension PCR. Primers C and D are complementary to different segments of the target DNA and contain sequences at their 5′ end which are complementary to each other. The deleted DNA segment is flanked by the 5′ end of each primer C and D that matches to the target sequences. Two different PCRs are performed in separate tubes, excess primers are removed from the PCR products, mixed and subjected to a PCR with primers A and B.

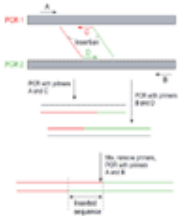

Figure 19. Insertion of small DNA segments by overlap extension PCR. Primers C and D are derived from the same target sequence but complementary to opposite strands. At their 5′ end sequences are added which are complementary to each other. The length of these sequences is limited by the oligonucleotide synthesis procedure. After PCR with primers A/C and B/D, performed in separate tubes, both reactions are mixed, excess primers are removed and a PCR with primers A and B is performed.

11.7. PCR for Analysis of Gene Expression

The detection of gene expression, i.e., synthesis of mRNA may be accomplished in several ways. Widespread methods for studying specific RNA molecules are in situ hybridization, Northern blots (⟶ Genetic Engineering – Nucleic Acid Hybridization), or RT-PCR.

In many cases the detection of a specific RNA molecule can be used to diagnose the existence of an infectious (Section Amplification) or genetic/cancerous disease. In both areas it is of importance to monitor the presence and the amount of viral RNA or cancerous mRNA indicating active infection or tumor development [35], [88]. With increasing knowledge of expressed gene sequences the expression patterns will reflect the biological status of cells with high precision yielding refinements in diagnostic strategies [89]. These applications underline the importance of quantitative (RT)-PCR technologies (Chap. Quantitative PCR).

[Top of Page]

References

General References

1. P. D. Bridge, D. K. Arora, C. A. Reddy, R. P. Elander: *Applications of PCR in Mycology*, Oxford University Press, Oxford – New York – Tokyo 1998.
2. Y. M. Dennis Lo, Rossa W. K. Chiu, Allen K. C. Chan (eds.): *Clinical applications of PCR*, Humana Press, Totowa 2006.
3. T. Dorak (ed.): *Real-time PCR*, Taylor & Francis, London, UK 2006
4. Th. Weissensteiner, H. G. Griffin, A. M. Griffin (eds.): *PCR Technology-Current Innovations*, CRC Press, Boca Raton-Ann Arbor 2003.
5. J. J. O'Leary, C. S. Herrington (eds.): *PCR 3: PCR In situ Hybridization — A Practical Approach*, IRL Press at Oxford University Press, Oxford – New York – Tokyo 1998.
6. M. A. Innis, D. H. Gelfand, J. J. Sninsky, T. J. White (eds.): *PCR Protocols — A Guide to Methods and Applications*, Academic Press Inc., San Diego 1990.

7. B. Kochanowski, U. Reischl (eds.): *Quantitative PCR protocols*, Humana Press, Totowa 1999.
8. D. S. Latchman (ed.): *PCR Applications in Pathology — Principles and Practice*, Oxford University Press, Oxford – New York – Tokyo 1995.
9. F. Lewis: *An Approach to in situ PCR,* PE Applied Biosystems, company brochure, Foster City 1996.
10. M. J. McPherson, P. Quirke, G. R. Taylor (eds.): *PCR — A practical approach*, IRL Press at Oxford University Press, Oxford – New York – Tokyo 2002.
11. M. J. McPherson, B. D. Hames, G. R. Taylor (eds.): *PCR 2 — A Practical Approach*, IRL Press at Oxford University Press, Oxford – New York – Tokyo 1995.
12. S. J. Meltzer: "PCR," in *Bioanalysis*, Humana Press, Totowa 1998.
13. C. R. Newton, A. Graham: *PCR*, BIOS Scientific Publishers, Oxford, UK 1997.
14. D. H. Persing (ed.): *PCR Protocols for Emerging Infectious Diseases*, ASM Press, Washington D.C. 1996.
15. Ch. Lovat, J. Logan, N. Saunders (eds.): *Real-Time PCR: An Essential Guide*, BIOS Scientific Publishers, Oxford, UK 2004

Specific References

16. R. K. Saiki *et al.*: "Enzymatic amplification of β-globin genomic sequence and restriction site analysis for diagnostic of sickle cell anemia," *Science* **230** (1985) 1350 – 1354. Links
17. K. B. Mullis, F. Faloona: "Specific synthesis of DNA in vitro by polymerase chain reaction," *Meth. Enzymol.* **155** (1987) 335 – 350. Links
18. D. Thiele: "The Technique of Polymerase Chain Reaction — a new diagnostic tool in Microbiology and other scientific fields (Review)," *Zbl. Bakt.* **273** (1990) 431 – 454. Links
19. R. K. Saiki *et al.*: "Primer-directed amplification enzymatic amplification of DNA with a thermostable DNA polymerase," *Science* **239** (1988) 487 – 491. Links
20. T. D. Brock: "Early days in yellowstone microbiology," *ASM News* **64** (1998) 137 – 140. Links
21. S. Cheng, C. Fockler, W. M. Barnes, R. Higuchi: "Effective amplification of long targets from cloned inserts and human genomic DNA," *Proc. Natl. Acad. Sci.* **91** (1994) 5695 – 5699. Links
22. M. R. Ponce, J. L. Micol: "PCR amplification of long DNA fragments," *Nucl. Acids Res.* **20** (1992) 623. Links
23. M. W. Berchtold, U. Hübscher: "PCR: basics and new developments," *Schweiz. Arch. Tierheilkd.* **138** (1996) 53 – 58. Links
24. G. R. Taylor, W. P. Logan: "Polymerase chain reaction: new variations on an old theme," *Curr. Opin. Biotechnol.* **6** (1995) 24 – 29. Links
25. I. A. Wilson: "Inhibition and Facilitation of Nucleic Acid Amplification," *Appl. Environ. Microbiol.* **63** (1997) 3741 – 3751. Links
26. A. Al-S. Waleed: "Optimisation of diagnostic PCR a study of PCR inhibitors in blood and sample pretreatment," PhD Thesis, Lund University, Faculty of Technology, Applied Microbiology, Lund March 24, 2000.
27. J. Sambrook, E. F. Fritsch, T. Maniatis: *Molecular cloning*, 2nd ed., Cold Spring Harbor Laboratory Press, New York 1989.
28. E. M. Southern: "Detection of specific sequences among DNA fragments separated by gel electrophoresis," *J. Mol. Biol.* **98** (1975) 503 – 517. Links
29. M. L. Metzker: "Emerging technologies in DNA sequencing," *Genome Res.* **15** (2005) 1767–1776. Links
30. F. Sanger, A. R. Coulson: "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerases," *J. Mol. Biol.* **94** (1975) 441 – 448. Links
31. F. Sanger, S. Nicklen, A. R. Coulson: "DNA sequencing with chain-terminating inhibitors," *Proc. Natl. Acad. Sci.* **74** (1977) 5463 – 5467. Links
32. R. Murali *et al.*: "Structural studies on an inhibitory antibody against Thermus aquaticus DNA polymerase suggest mode of inhibition," *Protein Eng.* **11** (1998) 79 – 89. Links
33. T. Moretti, B. Koons, B. Budowle: "Enhancement of PCR amplification yield and specificity using AmpliTaqGold DNA polymerase," *Biotechniques* **25** (1998) 716 – 722. Links
34. J. Wilhelm, A. Pingoud: "Real-time polymerase chain reaction," *Chembiochem.* **4** (2003) 1120–1128. Links
35. M. J. Espy, J. R. Uhl, L. M. Sloan, S. P. Buckwalter, M. F. Jones, E. A. Vetter, J. D. Yao, N. L. Wengenack, J. E. Rosenblatt, F. R. Cockerill 3rd, T. F. Smith: "Real-time PCR in clinical microbiology: applications for routine laboratory testing," *Clin. Microbiol. Rev.* **19** (2006) 165–256. Links
36. S. Yang, R.E. Rothman: "PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings," *Lancet Infect. Dis.* **4** (2004) 337–48. Links
37. V. S. Basleski: "The role of molecular diagnostics in the clinical microbiology laboratory," *Clin. Lab. Med.* **16** (1996) 49 – 60. Links
38. H. Shikata, N. Utsumi, H. Kuivaniemi, G. Tromp: "DNA-based diagnostics in the study of heritable and acquired disorders," *J. Lab. Clin. Med.* **125** (1995) 421 – 432. Links
39. A. Van Belkum, H. G. Niesters: "Nucleic acid amplification and related techniques in microbiological diagnostics and epidemiology," *Cell. Mol. Biol.* **41** (1995) 615 – 623. Links
40. B. Vogelstein, D. Gillespie: "Preparative and analytical purification of DNA from agarose," *Proc. Natl. Acad. Sci.* **76** (1979) 615 – 619. Links
41. R. J. Cano, H. N. Poinar: "Rapid isolation of DNA from fossil and museum specimens suitable for PCR," *Biotechniques* **15** (1993) 432 – 436. Links
42. Y. Itoh, N. Mizuki, T. Shimada, F. Azuma, M. Itakura, K. Kashiwase, E. Kikkawa, J. K. Kulski, M. Satake, H. Inoko: "High-throughput DNA typing of HLA-A, -B, -C, and -DRB1 loci by a PCR-SSOP-Luminex method in the Japanese

population," *Immunogenetics* **57** (2005) 717–729. Links

43. R. Boom *et al.*: "Rapid and simple methods for purification of nucleic acids," *J. Clin. Microbiol.* **28** (1990) 495 – 503. Links
44. Z. Li *et al.*: "Rapid detection of Mycobacterium avium in stool samples from AIDS patients by immunomagnetic PCR," *J. Clin. Microbiol.* **34** (1996) 1903 – 1907. Links
45. Y. Muramatsu *et al.*: "Improved method for preparation of samples for the polymerase chain reaction for detection of Coxiella burnetii in milk using immunomagnetic separation," *Vet. Microbiol.* **51** (1996) 179 – 185. Links
46. I. V. Wesley *et al.*: "Application of multiplex polymerase chain reaction for rapid identification of Campylobacter jejuni and C. coli associated with reproductive failure," *Am. J. Vet. Res.* **58** (1997) 1070 – 1075. Links
47. P. Markoulatos, N. Siafakas, M. Moncany: "Multiplex polymerase chain reaction: a practical approach," *J. Clin. Lab. Anal.* **16** (2002) 47–51. Links
48. J. B. Mahony: "Multiplex polymerase chain reaction for the diagnosis of sexually transmitted diseases," *Clin. Lab. Med.* **16** (1996) 61 – 71. Links
49. G. J. Nuovo: "Detection of viral infections by in situ PCR: theoretical considerations and possible value in diagnostic pathology," *J. Clin. Lab. Anal.* **10** (1996) 335 – 349. Links
50. K. E. Johansson, M. U. Heldtander, B. Pettersson: "Characterization of mycoplasmas by PCR and sequence analysis with universal 16S rRNA primers," *Methods Mol. Biol.* **104** (1998) 145 – 165. Links
51. S. H. Shah, T. L. Romick: "Subspecies differentiation of Salmonella by PCR-RFLP of the ribosomal operon using universal primers," *Lett. Appl. Microbiol.* **25** (1997) 54 – 57. Links
52. E. A. Wagar: "Defining the unknown: molecular methods for finding new microbes," *J. Clin. Lab. Anal.* **10** (1996) 331 – 334. Links
53. A. Sansila *et al.*: "Differentiation between Mycobacterium tuberculosis and Mycobacterium avium by amplification of the 16S-23S ribosomal DNA spacer," *J. Clin. Microbiol.* **36** (1998) 2399 – 2403. Links
54. L. F. Ward, J. C. Brown, G. P. Davey: "Two methods for the differentiation of Lactococcus lactis ssp. lactis and cremoris based on differences in the 16 S rRNA gene sequence," *FEMS Microbiol. Lett.* **166** (1998) 15 – 20. Links
55. K. P. Francis, G. S. Stewart: "Detection and speciation of bacteria through PCR using universal major cold-shock primer oligomer," *J. Ind. Microbiol. Biotechnol.* **19** (1997) 289 – 293. Links
56. S. Chow, K. Hazama: "Universal PCR primers for S7 ribosomal protein gene introns," *Mol. Ecol.* **7** (1998) 1255 – 1256. Links
57. M. Manzano *et al.*: "Development of a PCR microplate-capture hybridization method for simple, fast and sensitive detection of Salmonella serovars in food," *Mol. Cell. Probes.* **12** (1998) 227 – 234. Links
58. E. Fritz, D. Thiele, H. Willems, M. Wittenbrink: "Quantification of Coxiella burnetii by polymerase chain reaction (PCR) and a colorimetric microtiter plate hybridization assay (CMHA)," *Eur. J. Epidemol.* **11** (1995) 549 – 557. Links
59. R. C. McGlennen: "Dynamic mutations pose unique challenges for the molecular diagnostic laboratory," *Clin. Chem.* **42** (1996) 1582 – 1588. Links
60. D. Withcombe *et al.*: "A homogenous fluorescence assay for PCR amplicons: its application to real-time, single-tube genotyping," *Clin. Chem.* **44** (1998) 918 – 923. Links
61. R. Finke: "Theoretical basis and application of molecular diagnostics," *Exp. Clin. Endocrinol. Diabetes* **104** (1996) Suppl., 92 – 97. Links
62. R. P. Brown, J. Pestano: "Phylogeography of skinks (Chalcides) in the Canary Islands inferred from mitochondrial DNA sequences," *Mol. Ecol.* **7** (1998) 1183 – 1191. Links
63. V. Lucchini, E. Randi: "Mitochondrial DNA sequence variation and phylogeographical structure of rock partridge (Alectoris graeca) populations," *Heredity* **81** (1998) 528 – 536. Links
64. C. H. Lau *et al.*: "Genetic diversity of Asian water buffalo (Bubalo bubalis): mitochondrial DNA D-lop and cytochrome b sequence variation," *Anim. Genet.* **29** (1998) 253 – 264. Links
65. F. von Wintzingerode, U. B. Göbel, E. Stackebrandt: "Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis." *FEMS Microbiol. Rev.* **21** (1997) 213 – 229. Links
66. F. C. Tenover *et al.*: "Interpreting chromosomal DNA restriction patterns produced by Pulsed-Field Gel Electrophoresis: Criteria for bacterial strain typing," *J. Clin. Microbiol.* **33** (1995) 2233 – 2239. Links
67. A. Sander *et al.*: "Comparison of different DNA fingerprinting techniques for molecular typing of Bartonella henselae isolates," *J. Clin. Microbiol.* **36** (1998) 2973 – 2981. Links
68. S. Deng, C. Hiruki, J. A. Robertson, G. W. Stemke: "Detection by PCR and differentiation by restriction fragment length polymorphism of Acholeplasma, Spiroplasma, Mycoplasma, and ureaplasma, based upon 16rRNA genes," *PCR Methods Appl.* (1992) 202 – 204. Links
69. D. B. Goldstein, C. Schloterer (eds.): *Microsatellites: Evolution and Applications*, Oxford University Press, Oxford 1999.
70. P. H. Dear (ed.): *Genome mapping*, Oxford University Press, Oxford 1997.
71. J. Welsh, M. McClelland: "Fingerprinting genomes using PCR with arbitrary primers," *Nucl. Acids Res.* **18** (1990) 7213 – 7218. Links
72. J. Welsh *et al.*: "Genomic Fingerprinting by Arbitrarily Primed Polymerase Chain reaction resolves Borrelia burgdorferi into Three Distinct Phyletic Groups," *Int. J. Syst. Bacteriol.* **42** (1992) 370 – 376. Links
73. T. U. Westblom: "Molecular diagnosis of Helicobacter pylori," *Immunol. Invest.* **26** (1997) 163 – 174. Links
74. B. J. Bassam, S. Bentley: "DNA fingerprinting using arbitrary primer technology (APT): a tool or a torment," *Australas. Biotechnol.* **4** (1994) 232 – 236. Links
75. M. D. Burr, K. L. Josephson, I. L. Pepper: "An evaluation of ERIC PCR and AP PCR fingerprinting for discriminating Salmonella serotypes," *Lett. Appl. Microbiol.* **27** (1998) 24 – 30. Links

76. D. Moissenet *et al.*: “Comparative DNA analysis of Bordetella pertussis clinical isolates by pulsed- field gel electrophoresis, randomly amplified polymorphism DNA, and ERIC polymerase chain reaction,” *FEMS Microbiol. Lett.* **143** (1996) 127 – 132. Links
77. M. Gillings, M. Holley: “Repetitive element PCR fingerprinting (rep-PCR) using enterobacterial repetitive intergenic consensus (ERIC) primers is not necessarily directed at ERIC elements,” *Lett. Appl. Microbiol.* **25** (1997) 17 – 21. Links
78. A. H. McIntosh, J. J. Grasela, R. L. Matteri: “Identification of insect cell lines by DNA amplification fingerprint (DAF),” *Insect. Mol. Biol.* **5** (1996) 187 – 195. Links
79. J. Z. Jordens: “Characterisation of non-capsulate Haemophilus influenzae by repetitive extragenic palindromic (REP)-PCR,” *J. Med. Microbiol.* **47** (1998) 1031 – 1034. Links
80. J. E. Dawson *et al.*: “Fingerprinting of Ehrlichia species by repetitive element polymerase chain reaction,” *Am. J. Trop. Med. Hyg.* **57** (1997) 109 – 114. Links
81. L. R. Dice: “Measures of the amount of ecologic association between species,” *Ecology* **26** (1945) 297 – 302. Links
82. P. H. A. Sneath, P. R. Sokal: *Numerical taxonomy*, W. H. Freemann & Co., San Francisco 1973.
83. A. Stein, N. A. Saunders, A. G. Taylor, D. Raoult: “Phylogenic homogeneity of Coxiella burnetii strains as determined by 16S ribosomal RNA sequencing,” *FEMS Microbiol. Lett.* **113** (1993) 339 – 344. Links
84. T. Usichenko, A. Lattermann, G. Mueller-Eckhardt: “Serology versus ARMS-PCR in prospective HLA-class I typing for bone marrow transplantation,” *Ann. Hematol.* **77** (1998) 97 – 100. Links
85. C. A. MacCormick, H. G. Griffin, H. M. Underwood, M. J. Gasson: “Common DNA sequences with potential for detection of genetically manipulated organisms in food,” *J. Appl. Microbiol.* **84** (1998) 969 – 980. Links
86. A. Holst-Jensen, S. B. Ronning, A. Lovseth, K. G. Berdal: “PCR technology for screening and quantification of genetically modified organisms (GMOs),” *Anal. Bioanal. Chem.* **375** (2003) 985–993. Links
87. M. M. Ling, B. H. Robinson: “Approaches to DNA mutagenesis: an overview,” *Anal. Biochem.* **15** (1997) 157 – 178. Links
88. B. Kaltenboeck, C. Wang: “Advances in real-time PCR: application to clinical laboratory diagnostics,” *Adv. Clin. Chem.* **40** (2005) 219–59. Links
89. C. Wagner: “Molecular diagnostics,” *J. Mol. Med.* **75** (1997) 728 – 744. Links
90. PCR Jump Station: http://www.horizonpress.com/pcr/ This website contains informations and links to all aspects of PCR. Links to suppliers of reagents, commercial sequencing and DNA synthesis are available.
91. Comprehensive Protocol Collection: Molecular Biology Protocols: http://www.dartmouth.edu/~ambros/protocols/molbio.html Here you can find protocols for RNA preparation, plasmid, cosmid and phage DNA preparation, genome sequencing , southern blots, sequencing and PCR. Useful links to further techniques in molecular biology are provided.
92. National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov On this website you can search DNA and protein databases (GenBank, EMBL, SWISSPROT) for sequences and homologies or you can align and compare them by BLAST. Furthermore a free literature search program is available on this web site (PubMed MEDLINE)
93. REBASE: The restriction enzyme database: http://rebase.neb.com/rebase/index.html REBASE is a database of restriction enzymes, methylases and organsims producing these enzymes

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Reactive Distillation

MICHAEL SAKUTH, RWE-DEA AG, Herne, Federal Republic of Germany

DIETER REUSCH, Degussa-Hüls AG, Marl, Federal Republic of Germany

RALF JANOWSKY, Degussa-Hüls AG, Marl, Federal Republic of Germany

1. Introduction 2
2. Mathematical Modeling of Reactive Distillation Processes 3
2.1. Equilibrium-Based Models 3
2.2. Rate-Based Models 3
3. Design of Reactive Distillation Processes 4
3.1. Procedures for Process Design Studies 4
3.2. Flowsheet for Process Development . 7
4. Industrial Applications 7
4.1. Commercial Packing Structures 7
4.2. Industrial Catalytic Distillation Processes 10
5. References 12

Symbols and Abbreviations

a_j specific interfacial area, m^2/m^3
c_{tj} mixture molar densities, $kmol/m^3$
E energy transfer rate, W
${f_{ij}}^{\mathrm{L}}$ feed flow rate of component i to stage j in the liquid phase, kmol/s
k_j matrix of mass transfer coefficients, m/s
k_{ij} equilibrium ratio of component i on stage j
L_j liquid flow rate from stage j, kmol/s
L_{j-1} liquid flow rate to stage j, kmol/s
${N_{ij}}^{\mathrm{L}}$ liquid phase mass transfer rate of component i, kmol/s
N_j vector of mass transfer rates, kmol/s
N_{tj} total mass transfer rate, kmol/s
NC number of compounds
Q_{L} liquid phase heat loss, W
Q_{V} vapor phase heat loss, W
r_{ij} reaction rate of component i on stage j, kmol/s
${S_j}^{\mathrm{L}}$ ratio of liquid sidewithdrawal
V_j vapor flow rate from stage j, kmol/s
V_{j+1} vapor flow rate to stage j, kmol/s
X_i transformed liquid phase composition of component i
x_{ij} mole fraction of component i in liquid phase of stage j
${x_{ij}}^{\mathrm{I}}$ liquid mole fraction of component i in interface
x_j vector of liquid mole fractions
Y_i transformed vapor phase composition of component i
y_{ij} mole fraction of component i in vapor phase of stage j
${y_{ij}}^{\mathrm{I}}$ vapor mole fraction of component i in interface
y_j vector of vapor mole fractions
ΔH_r heat of reaction, J/kmol
ν_i stoichiometric coefficient for component i
ν_{T} sum of stoichiometric coefficients defined by Equation 9

Superscripts

I Interface
L liquid phase
V vapor phase

Subscripts

B bottom
D distillate
F feed
P product
R reference component
i component number
j stage number
t total

Abbreviations and Acronyms

AIBN Azo*bis*isobutyronitrile
CD Chemical distillation
DIPB Diisopropylbenzene
ETBE Ethyl *tert*-butyl ether

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.c22_c01

HB	High-boiling
HETP	Height equivalent to a theoretical plate
LB	Low-boiling
MB	Medium-boiling
MESH	Material balance/equilibrium condition/summation equation/heat balance
MTBE	Methyl *tert*-butyl ether
RD	Reactive distillation
RWD	Reaction with distillation
TAME	*tert*-Amyl methyl ether
TAEE	*tert*-Amyl ethyl ether
TIPB	Triisopropylbenzene

1. Introduction

Reactive distillation (RD) is a process in which a catalytic chemical reaction and distillation (fractionation of reactants) occur simultaneously in a single apparatus. From the reaction engineering view point, it can be classified as a two-phase countercurrent fixed-bed catalytic reactor.

This integrated reaction – separation technique is also known as catalytic distillation (CD) or reaction with distillation (RWD). According to [1], CD is a process in which a heterogeneous catalyst is localized in a distinct zone of a distillation column. RD is a more general term for this operation, which does not distinguish between homogeneously and heterogeneously catalyzed reactions in distillation columns. RWD is a trademark of the Koch Engineering Company for reactive distillation technology that uses their KataMax packing structures. This article exclusively deals with RD processes that operated with a heterogeneous catalyst system, i.e., CD technology.

The most important advantage of CD technology for equilibrium-controlled reactions is the elimination of equilibrium limitation of conversion by continuous removal of reactants or products from the reaction mixture. For example, unconverted reactants are removed in the vapor stream, while low-boiling products are fed downstream in the liquid phase, but the opposite case is also possible. It is the application of Le Chatelier′s principle to displace the chemical equilibrium by increasing the concentrations on the one side of the reaction and decreasing it on the other. The chemical composition at this equilibrium point can be calculated by means of the Gibbs energy of reaction at a given temperature. Activities must be used to recalculate the composition from the equilibrium constant, (i.e., the molar fractions of the components).

Usually, a partially converted reaction mixture, nearly in chemical equilibrium, leaves the fixed-bed reactor section and enters the CD column in the fractionating zone to ensure the separation of products from feedstock components. The fractionated unconverted feed components enter the catalytic section for additional conversion. The catalyst packing zone is installed in the upper or lower-middle portion of the column, with normal distillation sections above and below.

CD technology has several advantages over conventional operating methods, such as a fixed-bed reactor connected to a fractionating column, in which the distillate or bottoms are recycled after further separation steps. Apart from increased conversion, the following benefits can be obtained [1]:

- The most important benefit of CD technology is the lower capital investment, because two process steps can be carried out in the same device (process intensification). Such integration leads to lower costs for pumps, piping, and instrumentation.
- If CD is applied to an exothermic reaction, the reaction heat can be used to vaporize part of the surrounding liquid, which represents three fundamental advantages: The maximum temperature in the structured catalytic packing is limited to the boiling point of the reaction mixture, so that the danger of *hot spots* is reduced significantly. Also, extremely simple and reliable temperature control is achieved. In addition, the integration of reaction heat in the distillation process leads to energy savings by reducing reboiler duty.
- Product selectivities can be improved owing to fast removal of reactants or products from the reaction zone. Thus, the probability of consecutive reactions, which may occur in conventional operation, is lowered.
- If the reaction zone in the CD column is located above the feed point, poisoning of the catalyst can be avoided. This leads to longer catalyst lifetime compared to conventional plant.

There are three important constraints for applying CD technology to catalytic chemical reactions:

– The use of CD technology is only possible if the temperature window of the vapor – liquid equilibrium is equivalent to the reaction temperature. By changing the column operating pressure, this temperature window can be altered. The flexibility in temperature is not only restricted by the fact that two phases are required for disillation. Also the thermal stability of the catalyst can limit the upper operating temperature of the distillation column.
– Because of the necessity of wet catalyst pellets, the chemical reaction must take place entirely in the liquid phase.
– As it is very expensive to change the catalyst in the structured packing of a CD column, only catalysts with a long lifetime are suitable.

In the literature, it is also reported that endothermic reactions are not suitable for the CD technology, because it condenses vapor. Although endothermic reactions require more reboiler duty and therefore exhibit no large energy savings, there are no restrictions with regard to the application of the technology [2].

Chemical reactions which benefit from CD technology must fulfill the above-mentioned criteria. They include etherifications, esterifications, transesterifications, hydrations, hydrolysis, condensations, hydroisomerizations, oligomerizations, alkylations, and transalkylations.

2. Mathematical Modeling of Reactive Distillation Processes

2.1. Equilibrium-Based Models

Multicomponent separation processes, such as normal distillation processes, have been modeled by using the equilibrium-stage concept for a century. Therefore, early works on reactive distillation also used the *equilibrium-stage model* to simulate reactions with superimposed distillation.

The principal assumption of the equilibrium-stage model is that the vapor and the liquid stream that leave the stage are in thermodynamic equilibrium. In real distillation columns, of course, the residence time is too short to reach total equilibrium. For this reason, other efficiencies have been introduced into the model (e.g., Murphree efficiency, vaporization efficiency, etc.). MESH equations, which consists of:

– Material balance
– Equilibrium relations
– Summation equations
– Heat equations

are used to simulate conventional distillation columns. To introduce the chemical reaction further equations are needed to simulate reactive distillation processes (→ Reaction Columns, Chap. 2.). The simplest way to consider chemical reactions is to use an equilibrium constant, but many reactions are not fast enough to reach chemical equilibrium in one theoretical stage. Therefore, it is often necessary to use kinetic expressions, with depend on temperature and the concentration of each component.

2.2. Rate-Based Models

To describe the real phenomena of mass transfer in distillation processes in more detail, a second-generation nonequilibrium model was developed by Krishnamurthy and Taylor [3]. A nonequilibrium stage (Fig. 1) represents either a single tray or a section of packing in a column.

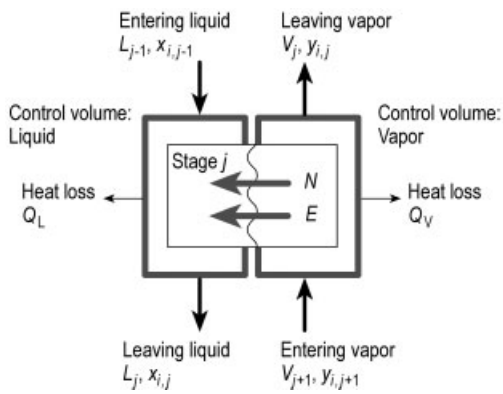

Figure 1. Nonequilibrium based stage model without feed and side streams

Contrary to the equilibrium-based stage model, this rate-based model treats the vapor and

liquid phases separately and combines them by means of a mass transfer rate and a heat transfer rate through the interfacial area. The total mass transfer rates are:

$$N_j^{\mathrm{V}} = c_{\mathrm{tj}}^{\mathrm{V}} k_j{}^{\mathrm{V}} a_j (y_j - y_j{}^{\mathrm{l}}) + N_{\mathrm{tj}}^{\mathrm{V}} y_j \quad (1)$$

$$N_j{}^{\mathrm{L}} = c_{\mathrm{tj}}^{\mathrm{L}} k_j{}^{\mathrm{L}} a_j (x_j{}^{\mathrm{l}} - x_j) + N_{\mathrm{tj}}^{\mathrm{L}} x_j \quad (2)$$

where $N_j{}^{\mathrm{V}}$ or $N_j{}^{\mathrm{L}}$ are the vectors of molar fluxes of the vapor or liquid stream, respectively, c_{tj} are the molar densities, k_j are the matrices of mass transfer coefficients, a_j is the specific interfacial area, and $N_{\mathrm{t}j}{}^{\mathrm{V}}$ or $N_{\mathrm{t}j}{}^{\mathrm{L}}$ are the total mass transfer rate of the vapor or liquid stream, respectively [4].

At the interface of each stage j phase equilibrium (E equation, see above):

$$K_{ij} x_{ij}{}^{\mathrm{l}} - y_{ij}{}^{\mathrm{l}} = 0 \quad (3)$$

and summation equations (S equations, see above) for each component i:

$$\sum_i y_{ij}^I - 1 = 0, \quad (4)$$

and

$$\sum_i x_{ij}^I - 1 = 0 \quad (5)$$

must be fulfilled.

To simulate the reactive distillation process, an expression for the reaction rate $r_{ij} = f$ (temperature, concentrations/activities) must be added. For catalytic distillation the chemical reaction is taken into account in the material balance of the bulk liquid phase

$$(1 + s_j{}^{\mathrm{L}}) L_j x_{i,j} x_{i,j-1} - L_{j-1} - f_{ij}{}^{\mathrm{L}} - N_{ij}{}^{\mathrm{L}} - r_{ij} = 0 \quad (6)$$

where $s_j{}^{\mathrm{L}}$ is the liquid side-withdrawal and $f_{ij}{}^{\mathrm{L}}$ is the feed flow rate of component i to stage j in the liquid phase. For many catalytic distillation processes, liquid – catalyst mass transfer must also be considered.

The reaction heat ΔH_{r} which is produced (exothermic reaction) or consumed (endothermic reaction) is considered in the heat balance of the liquid phase (see Figure 1).

Yuxiang and Xien [5] used this nonequilibrium stage model with simplified expression of the reaction kinetics based on concentrations to simulate the synthesis of MTBE, which they investigated experimentally in a column filled with catalyst granules in cloth pockets. The results of their simulation agreed fairly well with the experiments. However, Sundmacher and Hoffmann [6] investigated MTBE formation over a broad concentration range and found that the macrokinetics could be correctly described only by using liquid-phase activities for the intrinsic reaction kinetics and liquid – solid mass transfer.

The equilibrium-based stage model is an excellent *pragmatic approach* to simulate reactive distillation processes, especially those with homogeneous reactions. This model is part of all available process simulators (e.g., RADFRAC in ASPEN-PLUS, Aspen Technology, Inc.). For a *detailed study* of the superimposed distillation and interactions between reaction and mass transfer a rate-based approach is preferred (e.g., RATEFRAC in ASPEN-PLUS, Aspen Technology, Inc.).

3. Design of Reactive Distillation Processes

3.1. Procedures for Process Design Studies

Only a few systematic methods are available for the process design of multifunctional units such as reactive distillation columns. One method is the concept of reactive distillation lines with reactive azeotropes by Buzad and Doherty [7], Bessling et al. [8] and Stichlmaier et al. [9]. A brief description of this method is given below.

Figure 2 shows the transformation of coordinates for the equilibrium reaction $\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{C}$. This transformation reduces the equilibrium concentration of component C to a single point on the line of the transformed coordinates A and B. Mathematically one can define a chemical equilibrium reaction by Equation 7:

$$\nu_{\mathrm{a}} \mathrm{A} + \nu_{\mathrm{b}} \mathrm{B} \ldots \rightleftharpoons \nu_{\mathrm{p}} \mathrm{P} + \ldots, \quad (7)$$

where ν_i are the stoichiometric coefficients and P the desired products. The transformed liquid and vapor coordinates for an arbitrary component i, i.e., X_i and Y_i, are given by Equation 8, [10].

$$X_i = \frac{x_i - \frac{\nu_i x_\mathrm{R}}{\nu_\mathrm{R}}}{1 - \frac{\nu_\mathrm{T} x_\mathrm{R}}{\nu_\mathrm{R}}} \tag{8}$$

$$Y_i = \frac{y_i - \frac{\nu_i y_\mathrm{R}}{\nu_\mathrm{R}}}{1 - \frac{\nu_\mathrm{T} y_\mathrm{R}}{\nu_\mathrm{R}}}, \tag{9}$$

where R is a chosen reference component. R can be reactant or a product component. The stoichiometric coefficient ν_T in these equations is defined as:

$$\nu_\mathrm{T} = \sum_{i=1}^{\mathrm{NC}} \nu_i \tag{10}$$

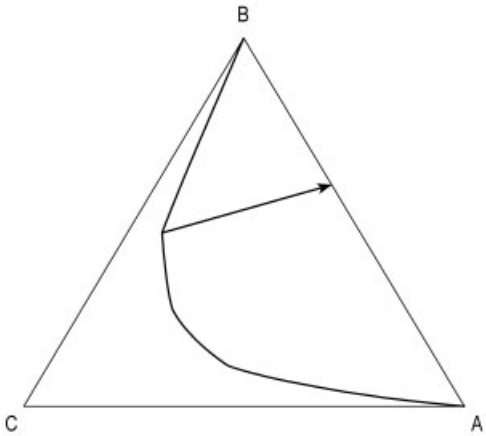

Figure 2. Transformation of coordinates for a equilibrium reaction A + B ⇌ C.

Both the vapor and liquid molar fractions of component P are set to zero, i.e., $X_\mathrm{P} = Y_\mathrm{P} = 0$. According to Equation (8), the transformed composition coordinates fulfill the unity constraints:

$$\sum_{i=1}^{\mathrm{NC}} X_i = \sum_{i=1}^{\mathrm{NC}} Y_i = 1 \tag{11}$$

By using these transformed coordinates for the feed components A and B, the mass balance and operating lines of a distillation column can be set up in the same way as for a system without a reaction. It reveals that feasible top and bottom compositions have to meet the lever rule (→ Distillation and Rectification, Chap. 4.2.2.) through the feed concentration and an appropriate distillation line, as shown in Figure 3. This is the same procedure as for conventional nonreactive systems.

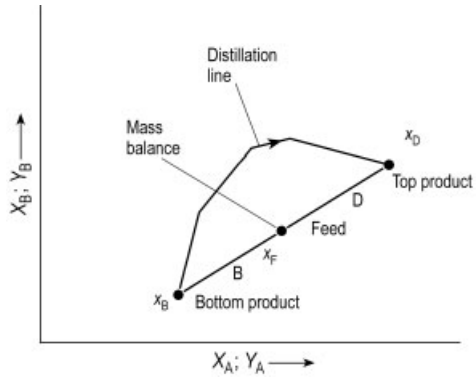

Figure 3. Application of the concept of reactive distillation lines. As mass balance must be fulfilled, the feed concentration X_F is located on a straight line between the transformed concentrations X_D and X_B.

Figure 4 shows reactive distillation lines for the equilibrium reaction A + B ⇌ C in the presence of an inert component D. Note that product C is no longer necessary, because of the transformed-concentration coordinates between A and B (see Eq. 8). For a given feed concentration F, two process layouts can be derived. In design type I, component C decomposes, and B can be separated from A and D. With design type II, component C is formed, and the inert component D is separated from a mixture of A and B, which is in chemical equilibrium with C. If the equilibrium constant of the reaction A + B ⇌ C is increased, a reactive azeotrope appears along line AB (Fig. 5). According to Ung and Doherty [11] this reactive azeotrope is characterized by identical vapor and liquid transformed-concentration coordinates (Eq. 12).

$$X_i = Y_i \tag{12}$$

In this case the reactive azeotrope and the inert component D are nodes; components A and B are saddles. Therefore, in contrast to Figure 4, it is not possible to realize design type I (decomposition of C). Only design type II, i.e., the formation of C and separation of D at the top and A, B, and C at the bottom of the RD column, is feasible. The concept of reactive distillation lines with reactive azeotropes can also be extended to several other reaction types and to nonideal reactive distillation systems as well.

In collaboration between the chemical industry and European universities, two tools

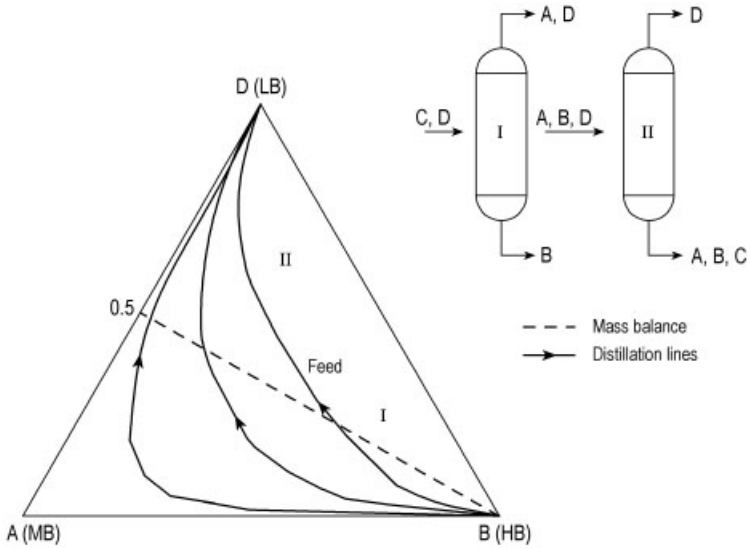

Figure 4. Reactive distillation lines for an equilibrium $A + B \rightleftharpoons C$ in the presence of an inert component D (LB = low-boiling, MB = medium-boiling, HB = high-boiling).

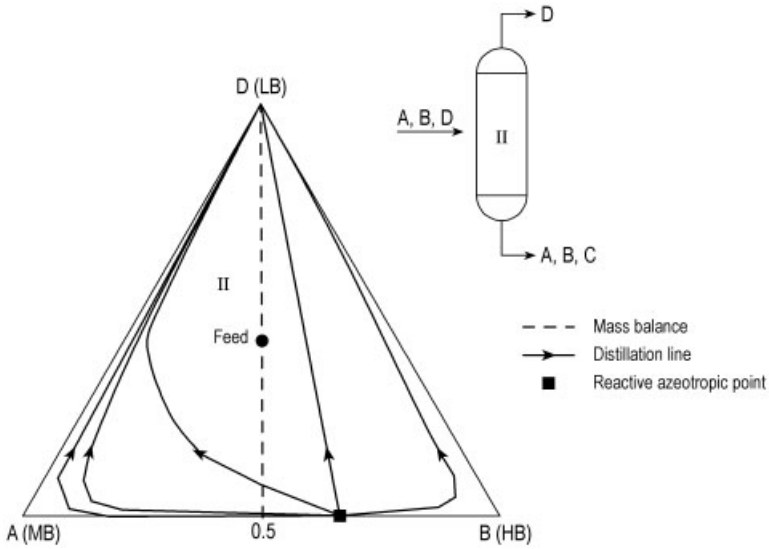

Figure 5. Reactive distillation lines for an equilibrium reaction $A + B \rightleftharpoons C$ in the presence of an inert component D with a reactive azeotropic point (LB = low-boiling, MB = medium-boiling, HB = high-boiling).

have been developed for the design of RD processes, namely SYNTHESIZER and DESIGNER. SYNTHESIZER is based on the above-mentioned analogy between normal distillation processes and RD processes. It investigates the feasibility of RD for an existing process during re-engineering or in the planning stage for a new process. DESIGNER gives a more precise insight into the RD process with respect to column parameters (e.g., column diameter, etc.). It contains calculation procedures with equilibrium-based stage models as well as rate-based models for a specific process design [12].

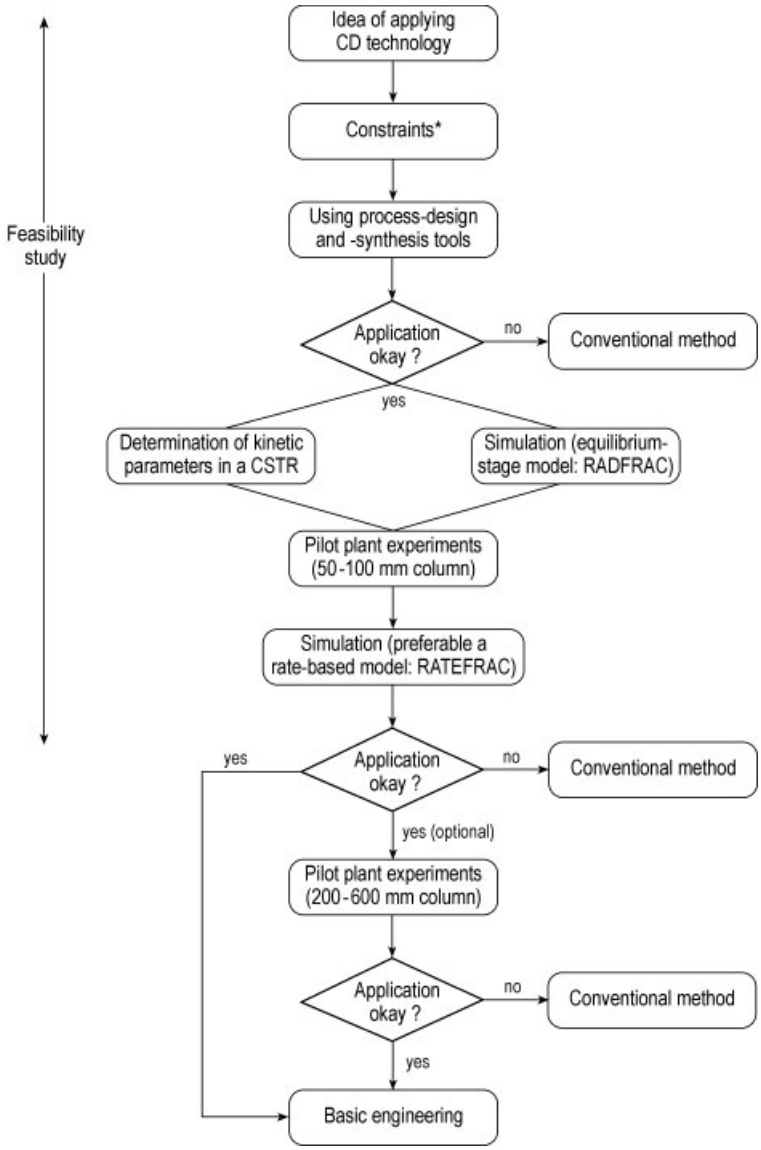

Figure 6. Flowsheet for the process development of a CD technology (CSTR = continous stirred tank reactor) (∗: see Chapter 1)

3.2. Flowsheet for Process Development

Figure 6 shows a flowsheet as a conceptual basis for the development of a CD process.

The key feature of this flowsheet is the feasibility study followed by pilot plant experiments (optional) and a basic engineering concept. After using the above-mentioned tools for process synthesis and design (see Section 3.1), the parameters of a kinetic model must be determined in an appropiate experimental setup.

Simulation studies with a simple equilibrium-based stage model like, such as RADFRAC (ASPEN-PLUS), should reveal the optimal feed point to the CD column, the operating conditions of the CD column, and the position of the reaction zone in the CD column. As a first approximation, these studies can be carried out by

using equilibrium constants to incorporate the chemical reaction in the column. For a detailed investigation, it is recommended to use an appropriate reaction expression with the determined kinetic parameters in a conventional simulator.

In pilot plant experiments, which are carried out in a CD column with an inner diameter of 50 – 100 mm, the simulation studies are examined. As the process is iterative, these experiments are needed to check and to improve the simulations.

As the last step of the feasibilty study, all obtained results are fed into a final simulation to decide whether an RD or conventional setup (e.g., reaction section followed by separation steps with internal recycle streams) should be applied. Preferably, a rate-based model such as for example RATEFRAC (ASPEN-PLUS), is used to check whether any mass-transfer limitations interfere with the removal of reactants. Such limitations can cause undesired consecutive reactions in the reaction zone.

4. Industrial Applications

4.1. Commercial Packing Structures

All commercial CD packing structures have in common that the catalyst is located in a distinct zone of a distillation column. Apart from the problem of positioning these packing structures, adequate contact between catalyst surface and liquid phase must be ensured.

Nowadays, in most industrial processes macroporous acidic ion-exchange resins are used, but any other solid catalyst pellet can be placed in the pockets of a structured packing. As the entire catalyst structure must be removed from the distillation column to exchange the catalyst, one evident prerequisite for the catalyst is a long lifetime (more than about three years). Catalyst replacement means shutting down whole plants which results in a significant loss of costs and operating time.

CD Tech Catalyst Bales. This type of spatial configuration in a distillation column was first developed and commercialized by CR & L (Chemical Research & Licensing Company) in 1980 [13]. It is now licensed by CD Tech (Chemical Distillation Technologies), a joint venture between CR & L and ABB Lummus Crest. In this technology the catalyst is supported in a fiberglass cloth, which is wrapped with stainless steel wire mesh. The demister wire mesh has two tasks. First, it stabilizes the fiberglass packing; second, it provides the void space necessary for distillation. These fiberglass cloths are rolled into catalyst bales and stacked on sieve trays in the column (Fig. 7).

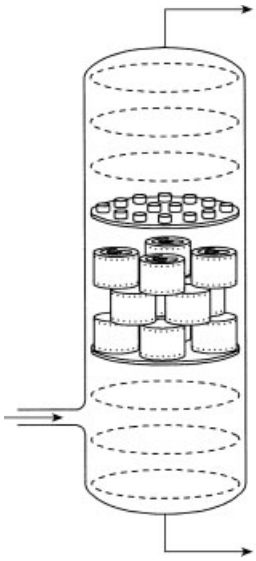

Figure 7. CD Tech catalyst bales

KataMax and KataPak Technology. Analogous catalytic packing configurations were developed in 1991/1992 independently by Koch Engineering Company as KataMax technology [14] and by Gebrüder Sulzer [15] as KataPak technology. Here the catalytic section in the CD column has a conventional corrugated structure, wire-cloth distillation packing which contains appropiate catalyst pellets for the desired reaction. The solid catalyst is held in an envelope of meshed and crimped screen, sealed at the edges. These envelopes are stacked together in modular blocks, so that the crimp of one envelope has a 90 degree angle of inclination to the crimp of the next. In this architecture, channels are formed for the rising vapor flow and descending liquid flow. At the intersections, they provide intensive mixing of vapor and liquid phases and radial distribution in a block. It is reported [16] that the performance

parameters of the KataMax packing (i.e., the hydraulic and mass transfer efficiencies) are similar to those of standard distillation devices (e.g., the Flexipac packing). At liquid loadings of $10-50\,m^3m^{-2}h^{-1}$ catalyst contact efficiency exceeded 75 – 90 %. Catalyst contact efficiency is the ratio of the rate constant of the catalyst in the packing to the rate constant of the catalyst in a stirred tank reactor for a simple liquid-phase first-order reaction at the same temperature.

Usually the catalyst content of KataMax or KataPak lies between 20 and 25 vol. %. For KataMax a HETP value (height equivalent to a theoretical plate) between 0.4 and 0.6 m can be used for simulation. For KataPak the HETP is 1.0 m at normal F factors (→ Distillation and Rectification, Chap. 8.2.1.). The specific surface area of KataMax is $210\,m^2/m^3$ and of KataPak, 85 or $125\,m^2/m^3$ (KataPak-S 170.Y or KataPak-S 250.Y, respectively).

The benefits of this type of packing are a very good distribution of liquid and vapor phases at a low pressure drop, efficient contact of the reactants with catalyst pellets and instantaneous distillative removal of reactants. Therefore, it combines excellent separation performance with an efficient mass and heat transfer for chemical reaction.

MultiPak Technology. This new structured catalyst packing was developed by Montz in 1998 in cooperation with the University of Dortmund [17]. In MultiPak technology the stacks are built up of alternating layers of meshed, but not crimped, catalyst envelopes and crimped structured distillation sheets. From the viewpoint of reaction and fractionating behavior, it has the same performance as KataMax or KataPak. The advantage of MultiPak is the identical geometric structure in an 80-mm laboratory packing and a 300- or 1000-mm pilot plant packing. Therefore, scaleup from pilot plant experiments to industrial plants is much easier. It is claimed that a defined flow pattern of vapor and liquid is achieved by this construction. This has advantages when resins are used as catalyst. As there is only liquid phase in the catalyst envelopes, the pellets are competely wetted under process conditions, and there is no swelling or shrinkage due to alternating contact with vapor and liquid. The disadvantage is a higher pressure drop along the structured packing.

Reactive Rings. The reactive rings of VEBA Oel [18] are based on packing bodies such as Raschig rings or Berl saddles. On the outer and inner surface of these bodies, macroporous ion exchange resins are bonded chemically or physically. For example, the active Raschig rings are prepared by using an open-celled sintered glass, which is impregnated with a mixture of styrene, a long chain alkane, *p*-divinylbenzene, and a radical starter (AIBN: azo*bis*isobutyronitril). After the polymerization procedure, it is sulfonated with chlorosulfonic acid to build up the catalytic performance.

This type of reactive rings have been used in an etherification process in a pilot plant column [6], [19].

Other Technologies. Another interesting, but as yet not commercialized, technology for immobilizing catalyst pellets in a distinct zone of a CD column is to incorporate a fixed-bed in the downcomer of a conventional tray column [20].

4.2. Industrial Catalytic Distillation Processes

Catalytic distillation can be used in catalytic chemical reactions which are limited by a chemical equilibrium. There are various reactions that satisfy this criterion, but only for etherification, esterification, and alkylation (synthesis of ethylbenzene or cumene) is this technology applied on an industrial scale. An overview of further possible applications can be found in [1], [19].

Etherification Processes (see also → Methyl *tert*-Butyl Ether). The application of CD technology to etherification is limited to synthesis of methyl *tert*-butyl ether (MTBE), ethyl *tert*-butyl ether (ETBE), *tert*-amyl methyl ether (TAME), and *tert*-amyl ethyl ether (TAEE).

Because of their good octane enhancing properties and low volatilities these ethers are excellent gasoline blending compounds (so-called oxygenates). Reduction of the tetraethyllead content in gasoline in the mid-1970s led to a dramatic increase in the demand for octane enhancers, with MTBE being used increasingly.

Driven by the U.S. Reformulated Gasoline Program, which mandated a minimum oxygen content in gasoline of 2 %, world MTBE production has now reached ca. 19×10^6 t/a [21]. The successful use of CD technology in MTBE production led to its worldwide establishment.

Table 1 summarizes processes for ether production based on CD technology. As shown in the simplified process flow diagram for the ETHERMAX process (Fig. 8), all processes have in common a reactor section followed by a CD column. There is no industrial process in which the whole conversion is performed in a single CD column. A partially converted mixture from the reactor section, which is nearly in chemical equilibrium, enters the CD column below the catalyst packing zone to ensure the separation of the ether from the feed stream. The catalyst packing is installed in the upper middle of the column, with normal distillation sections above and below.

Usually, the reaction of the isoalkene with methanol or ethanol is conducted in the presence of a slight stoichiometric excess of the alcohol. In addition to the advantages of a higher selectivity for the ether (lower formation of C_8 and C_{10} dimers) and shifting of the chemical equilibrium to the product side, this leads to a secure process control. In the absence of the alcohol in the reaction zone of the CD column exothermic dimerization and oligomerization of the C_4 and C_5 alkenes takes place at high reaction rates. A sharp and excessive temperature rise (hot spot) causes irreversible catalyst deactivation and catalyst damage. The excess alcohol is collected in the overhead of the CD column and separated from the hydrocarbon stream by extraction with water.

In the case of MTBE production, conversions of up to 99.9 % are possible by using this technology, if sufficient ether is present at the bottom of the column. Because of a minimum boiling azeotrope in the binary MTBE – methanol system, the ether carries the alcohol into the reaction zone of the CD column. This unusual vapor – liquid behavior and the high difference in the heat of vaporization between C_4 hydrocarbons and methanol are possibly responsible for the known multiple steady states in CD column operation [19], [22].

The TAME process usually gives lower conversions of 91 – 95 %.

Esterification Processes. Esterification is a good example for the beneficial use of the CD technology. In conventional operation the recovery of methyl acetate from the reactor outlet stream is complicated, because methanol forms a low boiling azeotrope with water and the separation of water from unconverted acetic acid is difficult. Due to these separation problems, the old Eastman Kodak process used a reactor coupled with eight distillation columns and an extraction column [24].

In the reactive distillation process, almost pure methyl acetate can be collected in the overhead of a single CD column at acetic acid conversions of greater than 99 %. The Eastman Kodak process uses sulfuric acid as catalyst [25] but a heterogeneous catalyst system (acid ion-exchange resin) can also be used successfully [26].

The conceptual basis for the successful implementation of reactive distillation in methyl acetate synthesis is shown in Figure 9. There are four zones in the column, which ensure the fairly high conversion. Acetic acid is separated from methyl acetate at the top of the column (zone I). In zone II, an extractive distillation section below the acetic acid feed extracts water from methyl acetate. The reaction takes place in the middle of the column (zone III). At the bottom (zone IV) methanol is fed and stripped from descending byproduct water.

Alkylation Processes. CD-technology is also applied for alkylation reactions in the case of ethylbenzene and cumene manufacture [1]. As the two processes are fairly similar, only the cumene process is discussed here.

In the CDCUMENE process, CD Tech catalyst bales are filled with an acidic, wide-pore zeolite catalyst, for example, zeolite Y, β or ω [27]. The bales are stacked in the middle of a distillation column. Below the reaction zone propene is fed; benzene is passed as reflux to the top of the column. The propene concentration in the reaction zone is held fairly low to slow down a side reaction in which diisopropylbenzene (DIPB) and triisopropylbenzene (TIPB) are formed. The cumene product with impurities of ethylbenzene and *n*-propylbenzene, is taken from the bottom of the column. The overhead pressure of the CD column is maintained at 5 bar.

Table 1. Overview of CD processes for ether production [23]

Process (licensor)	Process description	Catalyst	Market share∗
ETHERMAX (UOP)	fixed-bed tubular reactor or adiabatic recycle reactor followed by a reactive distillation column with *KataMax Packing*	sulfonated ion-exchange resin	ca. 32 %
CDMTBE & CDTAME (CD-Tech)	adiabatic reactor operating at the boiling point followed by a reactive distillation column with *CD Tech Bales*	acidic ion-exchange resin	ca. 58 %
CDETHEROL (CD-Tech)	adiabatic fixed-bed reactor operating at the boiling point followed by a reactive distillation column with *CD Tech Bales*	trifunctional ion-exchange resin	–
CATACOL (IFP)	recycle reactor, fixed-bed reactor followed by a reactive distillation column	acidic ion-exchange resin	ca. 10 %

∗ For MTBE, produced by CD technology in 1997

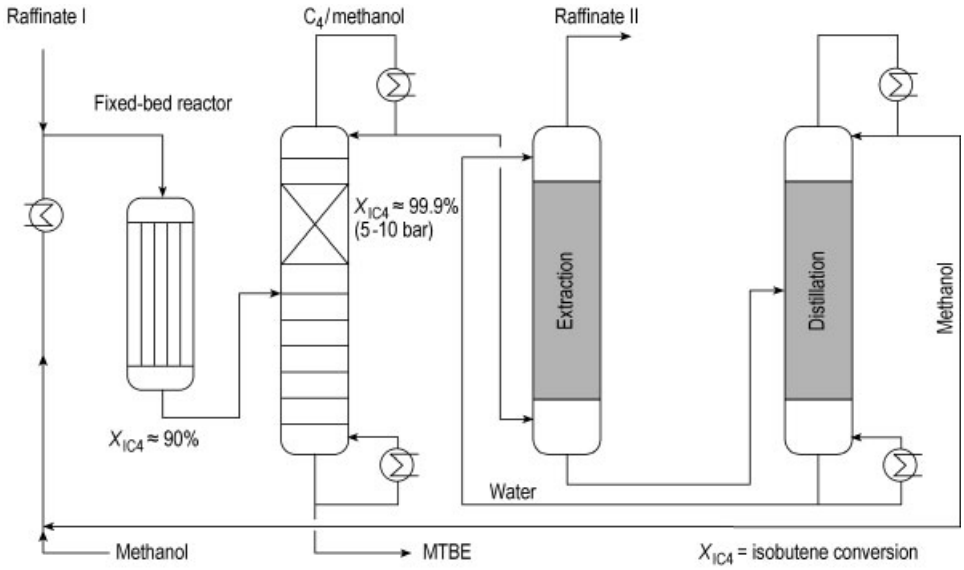

Figure 8. Simplified flow diagram for the ETHERMAX process

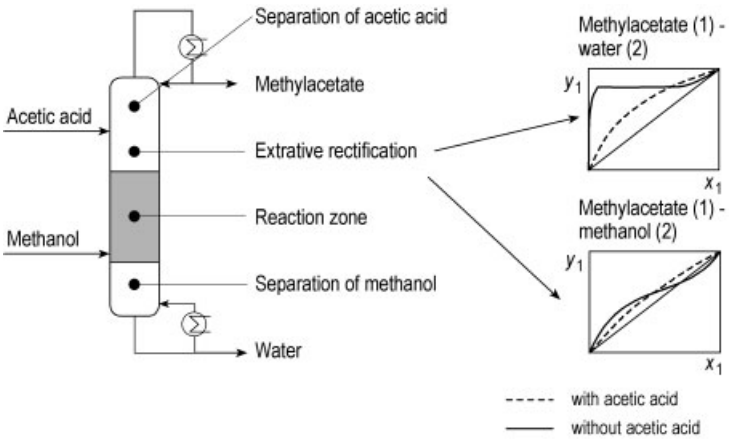

Figure 9. Methyl acetate synthesis with CD technology

The advantages of CD technology in cumene synthesis are lower formation of oligo-isopropylbenzenes (DIPB and TIPB), higher catalyst lifetime, and a higher conversion of benzene [28]. The internal benzene recycle to the reaction zone, which is a result of the distillative cumene – benzene separation in the lower portion of the column, ensures the observed higher conversion.

5. References

1. G. G. Podrebarac, F. T. T. Ng, G. L. Rempel, *CHEMTECH* **1997,** 37 – 45.
2. Hüls, EP 0 726 241, 1995 (M. Sakuth, U. Peters).
3. R. Krishnamurthy, R. Taylor, *AIChE. J.* **31** (1985) 449 – 456.
4. R. Taylor, R. Krishna: *Multicomponent Mass Transfer,* Wiley, New York 1993.
5. Z. Yuxiang, X. Xien, *Trans. IchemE* **70** (1992) 465 – 470.
6. K. Sundmacher, U. Hoffmann, *Chem. Eng. Sci.* **49** (1994) 3077 – 3089.
7. G. Buzard, M. F. Doherty, *Computers Chem. Engng.* **19** (1995) 395 – 408.
8. B. Bessling, G. Schembecker, K. H. Simmrock, *Ind. Eng. Chem. Res.* **36** (1997) 3032 – 3042.
9. J. Stichlmaier, T. Frey, *Chem. Ing. Tech.* **70** (1998) 1094 – 1095.
10. J. Espinosa, P. A. Aguirre, G. A. Perez, *Ind. Eng. Chem. Res.* **34** (1995) 853 – 861.
11. S. Ung, M. F. Doherty, *AIChE J.* **41** (1995) 2383 – 2392.
12. G. Schembecker et al., *Chem. Ing. Tech.* **70** (1998) 1096.
13. CR & L, US 4 232 177, 1980 (L. A. Smith Jr.).
14. Koch Engineering Company, EP 0 428 265, 1991 (A. P. Gelbein, M. Buchholz).
15. Gebrüder Sulzer, EP 0 396 650, 1992 (R. Shelden, J.-P. Stringaro).
16. J. L. DeGarmo, V. N. Parulekar, V. Pinjala, *Chem. Eng. Prog.* (1992), no. 3, 43 – 50.
17. Gebrauchsmuster No. 298 7 007.3, 1998, (A. Górak, L. U. Kreul).
18. VEBA Oel, US 5 244 929, 1993 (K. Gottlieb, W. Graf, K. Schaedlich, U. Hoffmann, A. Rehfinger, J. Flato).
19. K. Sundmacher, Dissertation, Universität Clausthal-Zellerfeld, 1995.
20. Koch Engineering Company, EP 0 664 721, 1997 (N. Yeoman, R. Pinaire, M. A. Ulowetz, T. P. Nace, D. A. Furse).
21. *Chem. Week,* Oct. 29, 1997, 44.
22. S. Hauan, T. Hertzberg, K. Lien, *Ind. Eng. Chem. Res.* **34** (1995) 987 – 991.
23. *Hydrocarbon Processing* **73** (1992) 104 – 110.
24. B. Bessling, Dissertation, Universität Dortmund, 1998.
25. V. H. Agreda, L. R. Partin, W. H. Heise, *Chem. Eng. Prog.* (1990) no. 2, 40 – 46.
26. J. Gmehling, J. Krafczyk, *Chem. Ing. Tech.* **66** (1994) 1372 – 1375.
27. G. R. Meima, M. J. M. van der Aalst, M. S. U. Samson, J. M. Garces, J. G. Lee, "Catalyst of Acid and Bases", *Proc. DGMK-Conf.* (1996) 125 – 137.
28. CR & L, US 4 849 569, 1989 (L. A. Smith Jr.).

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Water, Ultrapure

Standard Article

Dennis F. Hussey[1], Gary L. Foutch[2], M. Andy Ward[3]

[1]School of Chemical Engineering Oklahoma State University Stillwater, OK 74078, United States

[2]School of Chemical Engineering Oklahoma State University Stillwater, OK 74078, United States

[3]Olathe, KS 66062, United States

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.c28_c01
Article Online Posting Date: September 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Definition
2.1. Ionic Purity - Conductivity and Resistivity
2.2. Organic Purity
2.3. Particulate Purity
3. Technologies used for Ultrapure Water Processing
3.1. Ion Exchange Resins
3.1.1. Mixed-Bed Ion Exchangers
3.1.2. Homogeneous Beds
3.2. Membranes
3.2.1. Reverse Osmosis
3.2.2. Ultrafiltration and Nanofiltration
3.2.3. Electrodeionization
3.3. Adsorption - Activated Carbon Beds
3.4. Degasification
3.5. Clarifiers and Chemical Precipitators
3.6. Chemical Chlorine Removal
3.7. Ultraviolet Treatment
3.8. Membrane Contactors
3.9. Piping Materials, Gaskets
4. Analysis
4.1. Ion Chromatography
4.2. pH Measurement
4.3. Conductivity
4.4. Ion-Selective Electrodes
4.5. Determination of Total Dissolved Solids
4.6. Determination of Total Organic Carbon
4.7. Particle Counters
4.8. Colorimetric Analyzers for Silica and Phosphate
4.9. Dissolved Oxygen Electrodes
4.10. Capillary Electrophoresis
4.11. ORP Chlorine Analyzers
4.12. Nonvolatile Residue Monitors
4.13. Use of On-line Analyzers and Manually Collected Samples
5. Applications
5.1. Microelectronics Industry

5.2. Pharmaceutical Industry
5.3. Power Industry
5.3.1. Nuclear Power Plants
5.3.2. Fossil Fuel-Fired Plants
5.4. Chemical Industry
5.4.1. Boiler Feedwater Make-up
5.4.2. Process Water
6. Future Developments

[Top of Page]

1. Introduction

Water is an integral part of several “high-tech” manufacturing and processing industries. With severe processing requirements come stringent demands on all the raw materials and manufacturing components, included process water. Ultrapure water (UPW) is essentially a “high-tech” material as well. Consequently, significant technological efforts and expense are devoted to purifying water to severe standards.

Ultrapure water is produced in a series of separation units, the final step processing being primarily mixed-bed ion exchange (MBIE). MBIE is preceded by a range of stages where salts, biological contaminants, organic chemicals (usually collectively referred to as total organic carbon, or TOC), and dissolved gases are removed. These technologies include membrane processes (ultrafiltration, reverse osmosis RO, and pervaporation), adsorption typically on activated carbon for TOC and chlorine removal), disinfection e.g., by UV-light, degasification, or — most usually — a combination of these.

Industries that require ultrapure water include:

1. Power generation — Steam cycle water chemistry is critical to sustained operation of nuclear and coal-fired power plants. Ultrapure water processing occurs in three major areas of the plant production of make-up water, condensate polishing, and reactor water clean-up. Each of these applications has specific requirements for water quality and uses a range of different processes to achieve these requirements, including deep-bed ion exchange, powdered resin ion exchange, degasification and reverse osmosis.
2. Specialty materials processing — As the technology in the area of advanced materials becomes more and more complex, the purity of chemicals involved with their production and processing must also increase. Requirements for ultrapure water in the microchip and microelectronics industry are even more severe than for electrical power companies since contaminants, both ionic and particulate materials can effectively destroy microchip surfaces.
3. Biomedical and life science applications — The primary pharmaceutical applications for ultrapure water processing are: water for injection (WFI) and manufacturing of medicines. The separation requirements for pharmaceuticals are aimed at eliminating pyrogens (low molecular mass organic molecules — typical of membrane decay due to microbial decay that can cause fever. In addition, ultrapure water is used extensively as a diluent for dialysis. Another life science application of ion exchange includes water reuse on space missions.
4. Boiler feed water purification — ultrapure water is also required as boiler feed water in chemical plants, such as ethylene and ammonia production. High purity water is also needed by the pulp and paper plants.

[Top of Page]

2. Definition

Standards for ultrapure water vary among industries, but UPW is typically defined as having electrolytic conductivity less than 0.1 µS/cm. For comparison, the theoretical conductivity of pure water is 0.054 µS/cm at 25 °C. To be classified as ultrapure water, the ionic concentration must be less than 20 µg/L (industry often uses parts per billion, ppb NaCl or $CaCO_3$ equivalent). The NaCl/$CaCO_3$ equivalent of a specific ion is calculated by multiplying the actual concentration by the ratio of the ion equivalent weight and NaCl or $CaCO_3$ equivalent weights (58.5 and 50)

$$\mathrm{ppb}_{\mathrm{NaCl}} = \mathrm{ppb}_i \frac{\mathrm{EW}_{\mathrm{NaCl}}}{\mathrm{EW}_i}$$

$$\mathrm{ppb}_{\mathrm{CaCO_3}} = \mathrm{ppb}_i \frac{\mathrm{EW}_{\mathrm{CaCO_3}}}{\mathrm{EW}_i}$$

where EW is the equivalent weight and the index i denotes “ion”.

In microelectronics rinsing water — in particular, microchip manufacturing — sub parts per trillion (ppt, ng/L) concentrations of contaminants are achieved. For many semiconductor manufacturers, the desired contaminant concentrations are at, or below, the detection limits for most analytical equipment. On-line data are often confirmed with manually collected samples (grab samples) and external laboratory analysis. The operating temperature of the water is process dependent, although some processes require hot UPW (typically below boiling temperature), common applications require water between 22 – 28 °C. The allowable level of dissolved oxygen also depends on the process; some processes have no concentration limits, while others require no or small amounts of dissolved oxygen.

2.1. Ionic Purity — Conductivity and Resistivity

Ionic concentration limits in ultrapure water may be expressed for specific ions — where the concentration is monitored by on-line ion chromatography — or by the total conductivity of solution. In dealing with ultrapure water, resistivity, i.e., reciprocal conductivity, is easier to follow. Tables 1 and 2 list the water quality standards for ionic species in semiconductor grade water (Balazs Laboratories). Specifications change with system requirements and analytical capabilities, but the U.S.P., the ASTM, the Electrical Power Research Institute (EPRI), and other organizations post updates frequently. Some of this information is available on the World Wide Web (e.g., http://www.balazs.com and http://www.millipore.com).

Table 1. Semiconductor specifications for metals content (ppb (µg/L)) of ultrapure water by determined inductively coupled plasma-mass Spectrometry (ICP/MS) (Courtesy of Balazs Laboratories, 1993)

Metal	Attainable concentration	Acceptable concentration	Alert concentration	Critical concentration
Aluminum	0.007	0.05	> 0.0	0.20
Barium	0.002	0.01	> 0.05	0.10
Boron	0.3	< 2.0		
Chromium	0.008	0.03	> 0.03	0.05
Copper	0.005	0.05	> 0.05	> 0.20
Iron	0.01	0.10	0.20	> 0.20
Lithium	0.004	0.03	0.10	> 0.10
Magnesium	0.002	0.02	0.10	> 0.20
Manganese	0.004	0.03	> 0.03	0.10
Nickel	0.005	0.05	> 0.05	0.10
Sodium	0.01	0.06	> 0.20	> 0.50
Strontium	0.002	0.01	> 0.01	> 0.10
Zinc	0.008	0.06	> 0.05	> 0.10

Table 2. Semiconductor specifications for concentrations of ions in ultrapure water (mg/L determined) by ion chromatography (IC) (Courtesy of Balazs Laboratories, 1993)

Species	Attainable concentration	Acceptable concentration	Alert concentration	Critical concentration
Cations				
Na^+	< 0.01	0.05	> 0.02	> 0.5
K^+	< 0.02	< 0.1	> 0.02	> 0.5
NH_4^+	< 0.06	0.1	> 0.02	> 0.5
Ca^{2+}	< 0.02	< 0.1	> 0.01	> 0.2
Mg^{2+}	< 0.02	< 0.1	> 0.01	> 0.2
Anions				
F^-	< 0.1	< 0.1	> 0.02	–
Cl^-	< 0.02	0.1	> 0.02	> 0.5
Br^-	< 0.02	< 0.1	> 0.01	> 0.5
NO_3^-	< 0.02	< 0.1	> 0.01	> 0.5
HPO_4^{2-}	< 0.02	< 0.1	> 0.01	> 0.5
SO_4^{2-}	< 0.05	0.1	> 0.02	> 0.5

Power plant water chemistry has somewhat different UWP standards depending on the type of power plant, the water used to cool the condensers and the location within the plant. The primary objective of condensate polishing is to minimize erosion and corrosion. Elevated contaminant concentrations result in problems, due to chemical reactions, particularly in crevices, where water droplets concentrate these chemicals and allow rapid interaction with material surfaces. Routine sources of these contaminants are the materials of construction themselves and chemicals added to regulate water chemistry. Major problems may occur with leaks in condenser tube that add large amounts of contaminants rapidly to the system. Under these conditions, condensate polishers have the primary function of protecting the boiler and generator from these tube leaks. For nuclear power plants, EPRI has suggested water chemistry requirements for both pressurized-water-reactors (PWR) (Tables 3 and 4) and boiling-water reactors (BWR) (Tables 5 and 6) (see also ⟶ Nuclear Technology).

Table 3. EPRI PWR Reactor feedwater/condensate chemistry guidelines, December 1996*

Control parameter	Frequency of measurement	Action levels		
		1	2	3
Hydrazine, µg/L	daily	< 20	–	–
Total iron, µg/L	weekly	> 5.0	–	–
Total copper, µg/L	weekly	> 5.0	–	–
Dissolved oxygen, µg/L	continuously	< 15, > 200	–	–

* Action levels are defined by EPRI in the guidelines in detail. In short: Level 1: return to these limits as soon as possible; Level 2: take immediate action, reduce power to obtain these limits within 8 hours; Level 3: shut down as quickly as safety allows.

Table 4. EPRI PWR Reactor blowdown chemistry guidelines, December 1996

Control parameter	Frequency of measurement	Action levels		
		1	2	3
Cation conductivity (25 °C), µS/cm	continuous		> 1	> 4
Sodium, µg/L	continuous	> 5	> 50	> 250
Chloride, µg/L	daily	> 10	> 50	> 250
Sulfate, µg/L	daily	> 10	> 50	> 250

Table 5. EPRI PWR Reactor water chemistry guidelines, December 1996

Control parameter	Frequency of measurement	Action levels		
		1	2	3
Conductivity, µS/cm	continuously	> 0.3	> 1.0	> 5.0
Chloride, µg/L	daily	> 5	> 20	> 100
Sulfate, µg/L	daily	> 5	> 20	> 100
Zinc, µg/L		> 5	> 20	> 100

Table 6. EPRI PWR reactor feedwater/condensate chemistry guidelines, December 1996 (with permission)

Control parameter	Frequency of measurement	Action levels		
		1	2	3
Feedwater conductivity, µS/cm	continuously	> 0.065	–	–
Condensate conductivity, µS/cm	daily	> 0.1	–	–
Feedwater total copper, µg/L	weekly	> 0.5	–	–
Feedwater total iron, µg/L	weekly	> 0.5	–	–
Feedwater dissolved oxygen, µg/L	continuously	< 15, > 200	–	–
Condensate dissolved oxygen, µg/L	continuously	< 15, > 200	–	–

If action level is exceeded, analyze samples again within 96 h. If levels are not met, proceed with Action Level 1 responses.

Ionic concentration specifications are also established for other applications. Table 7 presents water quality specifications for the pharmaceutical industry.

Table 7. USP XXIII water specifications

Control parameter	Control limit

Conductivity, µS/cm	1.25
TOC, µg/L	500
Microbial contamination (cfu/mL) *	< 100

* cfu = colony forming units

2.2. Organic Purity

The U.S.P. establishes the water quality for pharmaceutical applications in the USA. Other countries have similar standards. The U.S.P. water quality standards focus on the presence of organic compounds, expressed as TOC. Organic content of water primarily originates from the degradation of biological waste matter (humic type molecule structure) or microbiological sub-fragments. Table 7 lists ultrapure water quality specifications given in U.S.P. XXIII. Table 8 gives specifications for organics content of water used in microelectronics manufacture.

Table 8. Semiconductor UPW specifications for final filter water quality

Test	Attainable value	Acceptable value	Alert value	Critical value
Resistivity (25 °C), MΩ/cm	18.2	18.2	17.9	17.5
TOC offline, µg/L	< 10	20	40	80
TOC online, µg/L	< 1	< 2	5	10
THM, µg/L	< 2	< 5		
Number of particle per liter determined by SEM *				
0.1 – 0.2 µm	< 200	< 1000	< 2000	
0.2 – 0.5 µm	< 100	< 500	< 500	
> 0.5 µm	< 1	< 50	< 100	
Number of particle per liter determined by laser online counter				
0.05 – 0.1 µm	< 100	< 500		
0.1 – 0.2 µm	< 50	< 300		
0.2 – 0.3 µm	< 20	< 50		
0.3 – 0.5 µm	< 10	< 20		
< 0.5 µm	< 1	< 4		
Bacteria count per 100 mL				
By culture: 100 mL	< 1	< 6	25	> 25
1000 mL sample	< 0.1	< 1		
by EPI	< 10	< 30	> 50	> 100
Silica, µg/L				
Dissolved	< 0.2	< 1	> 3	> 10
Total	< 0.5	3	> 5	> 10
Residue batch, ppm	< 0.1	0.1	> 0.1	> 0.3

* SEM = Scanning electron microscopy.

2.3. Particulate Purity

Particles in ultrapure water may come from several sources. These include microbes or their components, fibers (typically from upstream membranes used to make ultrapure water), nonmetallic particles (typically precipitates or resulting from breakdown of materials of construction within the processing loop), and metallic particles (iron oxides, etc.). Table 8 lists the particle count specifications for microelectronics grade water. Of particular importance are the detection requirements as a function of sample method (on-line or grab sample) and the analytical method used.

Particles on microelectronic surfaces can physically damage transistors or short-circuit the flow of electrons and effectively disable an entire circuit. Particle control in power plants is related to crud removal, which can otherwise interfere with proper heat transfer. For nuclear plants where water streams have low-level radioactivity, particles become the largest potential source of radiation exposure to employees.

[Top of Page]

3. Technologies used for Ultrapure Water Processing

3.1. Ion Exchange Resins [3], [10], [12], [14], [19], [20], [23], [27], [35-41], [43], [44]

Ion exchange is required to obtain ultrapure water at the most stringent purity conditions, i.e., a resistivity > 18 MΩ/cm. Process schemes that include all other technologies discussed in this article, but exclude ion exchange can only reduce ionic concentrations to the µg/L) level (5 – 9 MΩ/cm for reverse osmosis, 16 – 17 MΩ/cm for electrodeionization. As the resins become loaded with ions, they release ions back into solution in accordance with the law of mass action. Expressed for cation exchange resin in the hydrogen form and an arbitrary cation i:

$$K_{\mathrm{H}}^{\mathrm{i}} = \left(\frac{C_{\mathrm{H}}}{q_{\mathrm{H}}}\right)^{z_{\mathrm{i}}} \left(\frac{q_{\mathrm{i}}}{C_{\mathrm{i}}}\right)^{z_{\mathrm{H}}}$$

where $K_{\mathrm{H}}^{\mathrm{i}}$ is the selectivity of the resin for ion i with respect to hydrogen, C_H and C_i are the liquid phase concentrations of cations and protons, q_i and q_H are the resin phase concentrations, and z_i and z_H are the respective valence.

The use of ion exchange resins is also a two-edged sword, since at ng/L levels of impurities the resin itself becomes a primary source of contaminants. These contaminants include residual chemicals from resin regeneration — primarily sodium and chloride or sulfate ions — and degradation products from the ion exchange resins themselves — sulfonates from the cationic exchange resins and functional amines from anionic exchange resins. In addition, ion exchange resins are an expected source of low levels of particles and bacteria.

3.1.1. *Mixed-Bed Ion Exchangers (see also* ⟶ Ion Exchangers – Ion-Exchange Resin Combinations, ⟶ Ion Exchangers – Mixed Beds, ⟶ *Water – Demineralization) [42], [45-50]*

A particular advantage of ion exchange resins is that both cationic and anionic resins can be used together to purify water. Cationic resins release hydrogen ions while anionic resins release hydroxide ions. The hydrogen and hydroxide ions combine on into water molecules. Mixed bed ion exchange is used to process boiler feedwater, make-up water, steam-cycle condensate or for final-stage polishing (removal of trace contaminants). With regard to boiler feedwater, make-up water, and applications with less stringent UWP requirements, fixed-bed demineralizers followed by mixed-bed polishing units may be all the water treatment used.

Mixed-bed polishers are used in high pressure (> 3.55 MPa) feedwater applications, where coordinated phosphate or equilibrium phosphate chemistry is used as the boiler water treatment for pH control to prevent corrosion. In water used in coal-fired and nuclear power plants, microelectronic sites, and pharmaceutical applications a first stage MBIE produces pure water for additional water treatment.

The use of mixed-bed polishers accomplishes several goals. First, they usually reduce the total electrolyte concentration in make-up or boiler feedwater to nearly 0.04 to 0.09 mg/L and silica from 0.01 to 0.05 mg/L. Second, mixed-bed ion exchange reduces sodium and silica leakage that occurs at the beginning and end of each operating cycle to provide a more consistent, uniform feedwater quality. Third, the use of mixed-bed polishing provides protection against breakthrough of the primary demineralizer units. Mixed-bed polishers typically require infrequent regeneration; some units operate for several years without regeneration.

Mixed-bed demineralizers are usually monitored by observing the pH, conductivity, silica and sodium content.

The pH of the water exiting any primary mixed-bed demineralizer is typically 6.5 to 7.5, depending on the temperature and water chemistry. Whenever the pH drifts from normal operating conditions, the first action is to verify the accuracy of the sensor. If the sensor is functional, ionic breakthrough is occurring and the mixed bed polisher or fixed bed demineralizer should be regenerated.

Similarly, the conductivity of the mixed-bed unit may rise when breakthrough of the fixed-bed units occurs. When an increase of the product water conductivity or specific conductance occurs, the bed should be regenerated if needed.

Silica concentration in the effluent is a critical factor with mixed-bed demineralizers that provide boiler feedwater make-up to high pressure boilers because silica can volatilize at low concentrations in boiler water when operating at ≥ 10.4 MPa. The volatilized silica then deposits on turbine blades as the pressure is reduced across the turbine. Therefore, on-line monitoring of silica exiting the mixed bed is desirable when the product water is used for high-pressure boiler make-up.

Sodium analyzers are used to determine the breakthrough of sodium in mixed-bed demineralizers because sodium is usually the weakest held cation, and an increase in sodium effluent concentration indicates breakthrough is about to occur.

Regeneration. The separation of the cation and anion exchange resin is the most critical step in mixed-bed ion exchange regeneration. If the resins are separated poorly, some cationic beads are exposed to caustic and some anionic beads are exposed to acid. This cross-contamination results in higher equilibrium leakage — the initial baseline effluent concentration — and shorter service life.

The relative terminal settling velocities of the cation and anion exchange resins govern the separation of the resins. The terminal settling velocity of a spherical particle is directly related to its density and diameter, and the resins are manufactured to maximize the difference in settling velocity. Separation of the cation and anion exchange resin is accomplished with a backwash of water at a fluid velocity that is lower than the terminal settling velocity of the cation resin which is usually less than that of the anionic resin.

Conventionally manufactured ion exchange resin has a variable particle diameter and is classified by sifting the resin with fine screen meshes. Developments in resin manufacture have led to nearly uniform particle size, which enhances the

separation of the resin types. However resins with uniform particle diameter may lead to operational complications. The increased settling velocity difference may lead to resin stratification before being returned to service.

Regeneration of a mixed-bed vessel requires cation and anion resin separation followed by the injection of acid and caustic to the cation and anion resin, respectively. The amount of regenerant chemical is determined by the volume of resin to be regenerated.

Regeneration intervals vary with the influent water quality and flow rate, regeneration of resins used in boiler feedwater at an ethylene plant may occur in 12 – 24 h increments, PWR nuclear regeneration cycles may typically be 30 d, while a polish mixed bed in a semiconductor plant may last a year or longer.

Typically two types of regeneration are performed for mixed beds

1. In-situ regeneration (Fig. 1) — An inert resin with a terminal settling velocity between those of the cation and anion resins is used occasionally. Inert resin is used continuously during the service and regeneration cycles. It can foul and attract organics onto the surface by electrostatic charge. After the resin is separated in the vessel, caustic is added to the top of the vessel while acid is added simultaneously in the bottom. The regenerant streams mix in the inert resin layer and are discharged. Advantages of in-situ regeneration include reduced equipment and labor costs. However, a higher proportion of resin will be cross-contaminated, leading to higher initial contaminant concentrations at the start of the service cycle.

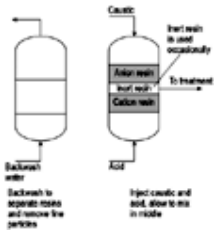

Figure 1. Process flow for in-situ regeneration

2. External regeneration (Fig. 2) — In external regeneration the mixed resin is backwashed and sluiced to a separate vessel, where it is backwashed until complete separation is achieved, then one of the resin layers is transferred to another vessel the other one remains in the separation vessel. Both resins are regenerated and then mixed, rinsed, and placed back into service, The water purity observed after initiation of the service cycle depends heavily on the efficiency of resin separation. Advantages of external regeneration include higher separation efficiencies of cation and anion resin, leading to lower equilibrium leakage. Disadvantages include high equipment and labor costs.

Figure 2. Example process flow for external regeneration

3.1.2. *Homogeneous Beds*

Homogeneous beds may contain several different types of resin, such as strong acid cation, weak acid cation, strong base anion, or weak base anion exchange resin. Demineralizers usually consist of cation units followed by anion units. The simplest and most common configuration is a strong acid cation exchange bed followed by a strong base anion exchange bed.

The function of the *strong acid cation exchange unit* is to remove the soluble cations (Ca^{2+}, Mg^{2+}, K^{+}, Na^{+}, Ba^{2+}, Sr^{2+}) from the feedwater stream. Divalent cations are more strongly held than monovalent cations. Na^{+} is often the first cation to leak through the strong acid cation exchanger because it is the weakest bonded species. As the cations are removed from the water, the resin releases hydrogen ions.

The strong *base anion unit* removes anions such as sulfate, chloride, alkalinity, and silica from the water and releases hydroxide ions into the water. Alkalinity is a measure of the buffering capacity of water. It is often associated with carbonic acid, but may include weak electrolytes. Sometimes it is defined as the concentration of OH^{-} ions. Silica is the weakest held anion, and it is usually the first one to leak from a strong base anion unit.

Two-bed systems with strong acid cation and strong base anion exchange resins typically generate water with a total electrolyte concentration of 2 to 3 mg/L, and a silica concentration of approximately 0.02 to 0.1 mg/L. A summary is given in Table 9 of homogeneous and mixed-bed service water flow rates in various applications.

Table 9. Service water flow rates for various homogeneous and mixed-bed ion exchange resin applications

Bed type	Flow rate		
	Minimum	Service	Maximum
Cation exchange bed	3.7 m/h	4.9 m/h	7.2 m/h
Anion exchange bed	3.7 m/h	4.9 – 12.2 m/h	13.5 m/h
Boiler feedwater mixed bed		4.9 – 490 m/h	
Condensate polishing mixed bed		120 m/h	
Microelectronics polish vessel		60 m/h	

Demineralizers produce their best water quality when they are operated continuously, because the ions stratify in the resin bed depending on their respective valences and the corresponding selectivities during first of the resin. When demineralizers are shut down and subsequently restarted, the demineralized water eluting first usually contains an increased amount of impurities because the ions leak from the resin to the water and gradually diffuse throughout the length of the bed. Therefore, an additional rinse step is required before the demineralizer is put back into service.

Regeneration. Demineralizer trains consisting of strong acid cation exchange beds followed by strong base anion exchange beds are usually regenerated simultaneously and separately. Table 10 summarizes the similarities and differences of cation and anion resin regeneration.

Table 10. Ion exchange resin regeneration specifications

Regeneration specification	Cation exchange resin	Anion exchange resin
Regenerant	sulfuric acid (U.S. nuclear) hydrochloric acid (European nuclear, some semiconductor	sodium hydroxide (occasionally potassium hydroxide for environmental concerns)
Basic procedure	1. backwash 2. 2 % acid injection (for sulfuric acid only) 3. 4 % acid injection (10 % for hydrochloric acid) 4. slow rinse (≈ 2 bed volumes) 5. fast rinse	1. backwash 2. 4 % sodium hydroxide chemical injection 3. slow rinse (≈ 2 bed volumes) 4. fast rinse
Regenerant dosage	80 – 160 kg/m^3 (98 % H_2SO_4) 80 – 400 kg/m^3 (30 % HCl)	32 – 160 kg/m^3 (process dependent)
Regenerant flow rate	0.6 m/h	0.6 m/h

First, the *anion resin* is backwashed using cation product water. The backwash removes fines and any particulate material that has been filtered by the cation unit and settled on the top of the anion resin. This backwash step minimizes channeling that can cause short cation run lengths. In addition, the bed is re-classified and fines (e.g., fractured resin beads) are removed. Second, the anion unit is regenerated using 4 % aqueous sodium hydroxide solution. The dilution water must not contain Mg^{2+} or Ca^{2+} to prevent precipitation of these cations in the resin bed induced by the high pH.

The exchange capacity increases with increasing sodium hydroxide dosage and exposure time (or flow rate). Once the specified amount of sodium hydroxide is added to the dilution water, one to two bed volumes of dilution water are passed through the bed at the same flow rate. This is called the slow rinse and its purpose is to allow the entire regenerant dosage to pass through the bed at the designed flow rate. Next, the anion unit is fast rinsed at about 3.7 m/h. The purpose of the fast rinse is to remove the excess sodium hydroxide from the anion resin.

The most common acids used for regeneration of strong *acid cation resin* are sulfuric and hydrochloric acid. When sulfuric acid is employed as the regenerant, it is important to use a stepwise regenerant dosage, especially in water that contains a significant concentration of calcium ions to prevent calcium sulfate precipitation in the resin bed. When hydrochloric acid is used as the regenerant, a 10 % concentration is often maintained throughout the regeneration, precipitation of calcium

salts not occur.

3.2. Membranes (⟶ Water – Membrane Separation Processes in Water Treatment) [53-58]

Several membrane technologies are used in ultrapure water processing; reverse osmosis, nanofiltration, and ultrafiltration. The choice of membrane type depends on inlet water quality and required throughput. In general, the larger pore size ultrafiltration is used upstream in the process while RO is used to further treat processed water. Ultrafiltration helps minimize fouling of the more restrictive RO pores by not exposing them to a higher level of contaminants. If the feedwater does not produce excessive fouling of the RO, it is possible to use RO as an initial purification stage.

Nonporous membranes are available for specialized applications, such as pervaporation (⟶ Membranes and Membrane Separation Processes – Principle of Pervaporation). Water “dissolves” into the membrane structure and passes through by concentration flux. A vacuum on the permeate stream generates the concentration gradient.

3.2.1. *Reverse Osmosis [62-67]*

Several pressures and flow rates are normally monitored in a reverse osmosis system, and they are summarized in Table 11. In addition to these pressures and temperatures, the feed water temperature, the silt density index (SDI) and product and reject conductivities are monitored.

Perhaps the three most useful parameters for monitoring reverse osmosis operation are the normalized flow, percent salt rejection, and pressure drop (either the pressure drop across the individual element or that across the pressure vessel).

Table 11. Pressures and flow rates that should be monitored in a reverse osmosis system

Parameter	Location of sensor
Pressures	Pressure in cartridge filter
	Pressure after cartridge filter
	Pressure inlet into booster pump
	Feedwater booster pump discharge pressure
	Inlet feedwater pressure to each stage
	Brine (concentrate) pressure
	Permeate (product) pressure
Flow rates	Total feedwater flow rate
	Feedwater flow to each stage
	Permeate flow from each stage
	Total permeate flow
	Total brine flow

Normalized flow is used to monitor flow performance while correcting for change in water temperature. More water can pass through a reverse osmosis membrane when the water temperature is elevated because the pore size increases with increasing temperature.

Salt rejection is defined as the salt concentration in the feedwater minus the permeate concentration divided by the concentration in the feedwater. Salt rejection is usually expressed as a percentage.

The *pressure drop* per vessel should be limited to 500 kPa, and the pressure drop per element to 200 kPa. Excessive pressure drop can cause the membrane to compact and cause a permanent loss of flow. Periodic cleaning of the reverse osmosis membranes, or elements, should be done when the pressure drop begins to approach the limitations.

Reverse osmosis systems require cleaning when the normalized flow is reduced by 15 % or when the pressure drop per vessel increases to near the manufacturer's limits. This impaired performance can be due to particulate fouling, scale, organic fouling, and microbiological contamination.

Particulate fouling can occur when the pretreatment system does not remove the suspended solids present in the feedwater to the RO unit. Examples of particulate fouling include silt, particulate iron, and other corrosion products.

Scale occurs when salts exceed their solubility limit. This can occur as a result of many causes, including malfunction of pH control equipment a too high percent recovery of the RO unit, a too low addition or lose of antiscalant, or use of the wrong antiscalant.

Organic fouling occurs on RO membranes because the membrane surfaces are organic materials and have an affinity for other organic materials. Examples of common organic foulants include tannic acid and humic acid. Microbiological contamination can be either the slime of living organisms or the remains of dead organisms.

3.2.2. *Ultrafiltration and Nanofiltration [68-70], [81-88]*

Ultrafiltration and nanofiltration are similar in some respects to RO. The pore size of ultrafiltration and nanofiltration membranes is much larger, and they allow much higher flow rates. Both processes can be used effectively with UPW

applications as preliminary separation equipment. Nanofiltration can remove larger molecule, particulate, and divalent species: all but simple elemental ions and low molecular mass organics. The cutoff size is about 0.001 µm, compared to 0.01 µm for ultrafiltration and 0.0001 µm for RO. Both technologies can play a significant role in upstream processing by reducing the load on downstream equipment. Nanofiltration may reduce the hardness in the feedwater supply.

Pressure drop is the primary monitoring variable for adequate functioning of membranes. Increasing pressure drop is a sign that the membrane is accumulating mass on the surface. In some cases downstream water samples are filtered and examined by microscope. The presence of fibers is an indication of potential membrane failure. Exposure to ppm levels of oxidants, such as chlorine, can degrade cellulose acetate membranes to the point that they fail to meet UWP standards.

3.2.3. *Electrodeionization [71-79]*

Electrodeionization (EDI) is a relatively new separation technology that has found commercial application since the mid 1980s. EDI combines mixed-bed ion exchange with electrochemistry to split a feed stream into a product and concentrate stream.

As illustrated in Figure 3, the EDI unit works by placing a mixed-bed ion exchange resin between two special membranes that allow the migration of only cations or anions. Several mixed bed resin/membrane layers are placed between an anode and cathode, a direct current is applied between both electrodes. The influent is then passed through the mixed bed, where the cations migrate to the cathode through the cation membrane into the cation concentrate stream, while the anions migrate to the anode through the anion membrane into the anion concentrate stream. The concentrates are collected and led to the discharge effluent. Water is also electrolytically spit into H^+ and OH^- ions, thus providing a simultaneous regeneration mechanism for the cation and anion resin that does not require chemicals.

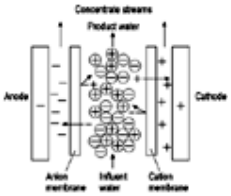

Figure 3. Electrodeionization unit configuration

Electrodeionization units have several advantages; such as continuous operation without regeneration, no regeneration chemicals, and relatively small power requirements. Disadvantages of EDI units were initially high capital costs, mechanical/electrical complications, low flow rates, and susceptibility to fouling. Current EDI technology has equalized the capital costs to that of a mixed-bed, and flow capacity has been increased to be comparable to mixed-bed ion exchange columns for some applications.

3.3. Adsorption — Activated Carbon Beds

Granular activated carbon beds can be used to remove chlorine from water before it enters either demineralizers or reverse osmosis systems. Chlorine removal by reaction with activated carbon is as follows [94]:

$$2\,Cl_2 + C + 2\,H_2O \rightarrow 4\,HCl + CO_2$$

6 kg of chlorine consume 1 kg of carbon. However, the use of activated carbon to remove chlorine from water streams is limited by the short life of carbon beds due to fouling. In addition, a carbon bed may be a microbiological growth site.

3.4. Degasification [95-97], [99]

Two types of degasification are prevalent in industry, forced draft and vacuum degasification.

Forced Draft Degasification (Decarbonator). Forced draft degasification is often used to remove carbon dioxide that may be physically dissolved in the water after the pH drop following a homogeneous cation exchange bed. A decarbonator is a column that has plastic packings at the top and a clear well underneath. The process water is spayed over the plastic packings, which increases the air – water interfacial area, and a low-pressure fan blows air from the bottom. The air strips the water of carbon dioxide.

Vacuum Degasification is used for the removal of dissolved oxygen and other gases. Process water is introduced into a tower that is under vacuum. The water approaches equilibrium with the low-pressure environment, liberating most dissolved gases into the vapor phase. The removal of dissolved oxygen is further advanced by the addition of pure nitrogen. The higher partial pressure of nitrogen leads to a lower partial pressure of oxygen and increases the driving force for oxygen to leave the liquid phase and enter the vapor phase.

3.5. Clarifiers and Chemical Precipitators

Clarifiers and chemical precipitation units are commonly used to pretreat water before its use as feedwater for demineralizers or RO.

Clarifiers are used to remove silt and other suspended solids from water. Cationic coagulants, sometimes used in conjunction with anion flocculants, are used to neutralize the silt and colloidal material present in many water supplies. Other clarifiers use coagulants such as alum to remove the suspended solids and adsorb the colloidal material.

Chemical precipitation units generally use lime in conjunction with organic or inorganic coagulants to remove suspended solids, colloidal material, and some of the dissolved solids in the water supply.

3.6. Chemical Chlorine Removal

Chlorine must be removed from feedwater before it enters ion exchange or RO systems to prevent oxidation of the ion exchange resin or RO membranes. Sodium sulfite, sodium bisulfite, sulfur dioxide, or calcium sulfite are among the chemicals that can be used for chlorine removal.

Sodium sulfite and sodium bisulfite react with chlorine according to:

$$Na_2SO_3 + Cl_2 + H_2O \rightarrow Na_2SO_4 + 2\,HCl$$

$$2\,NaHSO_3 + 2\,Cl_2 + 2\,H_2O \rightarrow Na_2SO_4 + 4\,HCl + H_2SO_4$$

Calcium sulfite is a granular material that slowly dissolves as it reacts with chlorine. The reaction is shown below:

$$CaSO_3 + Cl_2 + H_2O \rightarrow CaSO_4 + 2\,HCl$$

3.7. Ultraviolet Treatment [89-93]

Ultraviolet lamps serve two distinct purposes in ultrapure water manufacture; oxidizing organic compounds and sterilizing microbial contaminants. Oxidation of organic compounds requires 185 nm UV light in the presence of an oxidizing agent such as ozone or hydrogen peroxide. The compounds are oxidized to carbon dioxide and water. The carbon dioxide and remaining oxidizing agent may be removed downstream in a degasification unit. Microbial sterilization is typically accomplished with 254 nm UV radiation. The dead cell matter is removed downstream with a cartridge filter.

Some advantages of UV lamps include they are inexpensive, have relatively low maintenance requirements, and they destroy organic compounds instead of transferring them to another phase. The primary disadvantage of UV lights is their application is limited to pure water because turbidity and particulate matter affects the performance of the units.

3.8. Membrane Contactors [103-105]

Membrane contactors are useful for many gas – liquid and liquid – liquid separations. The principle of operation is analogous to that of a heat exchanger, except the primary driving force is the concentration difference between two fluids instead of temperature difference. Contaminants diffuse through a porous membrane from fluid A where they are present in high concentrations to another fluid B where the contaminant is not present.

Membrane contactors may be used instead of vacuum degasification towers for the removal of dissolved oxygen and other gases from ultrapure water. Figure 4 describes the mass transfer mechanism for transferring oxygen from water to the sweep gas. The contactor is typically cylindrical in shape, and has two inlets and outlets for the sweep gas and water. The sweep gas is typically pure nitrogen that may be under vacuum. Some applications require no dissolved nitrogen, in which case a vacuum is applied to the gas side of the membrane.

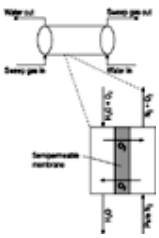

Figure 4. Mass transfer mechanism for membrane contactor

The advantages of membrane contactors include reduced space requirements and no mechanical parts. The disadvantages of membrane contactors include the possibility of particulate and organic fouling, cost, and maintenance of vacuum pumps if required.

3.9. Piping Materials, Gaskets [107-113]

Piping materials for ultrapure water vary upon the application. The nuclear industry uses stainless steel in the demineralizer make-up plant, where concentrations of contaminants are highest, and carbon steel in polisher vessels. The polishers tend to be rubber-lined, due to potential contact with regeneration chemicals. The semiconductor industry must use specialty plastics that do not shed particles and have low TOC leachables. Leachable TOC compounds include solvents or other organics within the polymer matrix that diffuse into solution. There are several plastics that may be used, such as poly(vinylidene fluoride) (PVDF), perfluoroalkoxypolymer (PFA) and poly(tetrafluoroethylene) (PFTE). Gasket materials must not generate particles; some common materials include teflon and silicone.

[Top of Page]

4. Analysis

4.1. Ion Chromatography [126-136]

Ion chromatography (IC) uses packed columns of ion exchange material to separate ionic species in a sample. As the separated peaks pass a conductivity detector, their relative concentrations are determined, as compared to the standard eluant concentration. With concentrator capabilities, IC can be used to detect ionic concentrations less than 1 ng/L.

Ion chromatography is the best method to determine a wide range of specific ion concentrations. IC can be used for either batch sampling or can be set up on-line. On-line IC typically reports lower readings than manually collected (grab)

samples. On-line sampling is only done when the severest specifications are required because of the expense and labor requirements of the IC. Even then, plants may forgo on-line IC analysis for other monitoring techniques supported by occasional IC grab samples.

4.2. pH Measurement

Monitoring the pH of ultrapure water is commonly done throughout the treatment plants, although at near neutral pH its value is somewhat limited. Although precision pH devices can measure pH to ± 0.01 this is not sufficient to represent accurate concentrations at ng/L levels. This can be seen by considering that for theoretically perfect water at pH of 7.00 0.1 µg/L H^+ and 1.7 µg/L OH^- present. However, UPW contaminant concentrations of interest are three to four orders of magnitude lower.

pH measurement is most applicable for power plants that add amines for corrosion control, for make-up water processing, for regeneration and rinse cycles, as an indicator of MBIE breakthrough, and plant concentration excursion incidents. However, since pH is such a basic water property and pH measurement relatively inexpensive, it is found on UPW lines as a normal monitoring tool.

4.3. Conductivity

Perhaps the most common method for determining the total ionic contaminants concentration in water is measuring the electrical conductivity.

At low concentrations, the conductance of water increases linearly with increasing concentration of ionic species.

Conductivity is temperature dependent, but most sensors have a temperature compensation method programmed in the electronics. The primary disadvantages of conductivity measurements are that only the total ionic concentration is estimated, specific ionic species are not measured, and conductivity sensors are generally ineffective detecting single digit ng/L ionic concentrations.

4.4. Ion-Selective Electrodes

Ion-selective electrodes can be used effectively in feed water and make-up water of demineralizers and RO units. However, for ultrapure flow streams ionic concentrations are below the detection limits of these devices currently. In addition, even at detectable parts-per-million ranges, some ionic species interfere with each other so that an accurate concentration of a specific ion may not be possible. A notable exception are sodium analyzers; current on-line sodium monitors are based upon a glass electrode that responds selectively to sodium. The hydrogen ion interference is minimized with the addition of ammonia or an organic amine such as diisopropylamine. Detection limits of 10 ng/L have been claimed; however, in practice the nominal detection limit varies from 50 – 100 ng/L.

4.5. Determination of Total Dissolved Solids

The quantity of total dissolved solids (TDS) is a measure of all solids, ionic or molecular, which remain after a sample is evaporated. This measurement is nonspecific in nature, and generally of advantage for the higher concentration streams within the UPW plant.

4.6. Determination of Total Organic Carbon [137-146]

A variety of total organic carbon (TOC), sometimes referred to as total oxidizable carbon analyzers are available in industry. Low-level analyzers operate by precisely recording the temperature and conductivity of a sample, oxidizing the sample to convert the organic compounds to carbon dioxide and water, and then measuring the conductivity of the sample again. The oxidation of simple hydrocarbons is described by the following generalized equation.

$$C_xH_y + [xy/4]O_2 \rightarrow xCO_2 + [y/4]H_2O$$

The change in conductivity is compared to a known calibration curve and the measurement is recorded (typically in ng C/g solution). The oxidation technique varies with the instrument; typically it is either a wet-chemistry or photocatalytic oxidation.

4.7. Particle Counters [25], [147-153]

Liquid particle counters operate by passing a small sample through a flow cell while a laser is focused on a photometric detector. The photometric detector sends a current proportional to the intensity of the laser to a recording device. When a particle passes through the flow cell, the particle scatters the laser light. The reduction of the measured intensity on the photometric detector is proportional to the effective diameter of the particle. The reduction in current is counted as a single particle, with a diameter determined by the change in the current.

The number of particle counts per given sample volume is the typical unit of measurement. Significant advances in on-line particle counters include increased sample flow rate and a particle diameter detection limit of 0.05 µm.

4.8. Colorimetric Analyzers for Silica and Phosphate [31], [32], [91], [92]

Silica detection is difficult with standard ion measurement techniques because silica is present in ionic, molecular, and particulate forms. The amount of ionic silica can be measured using a silicomolybdate/heteropoly acid blue colorimetric method. A sample of known volume is introduced in a cell and dosed with a molybdate reagent that reacts with available silica and phosphate to form molybdosilicic and molybdophosphoric acids. After several minutes, another acid (typically sulfuric acid or a proprietary mixture) is added to consume the monophosphoric acid. A photometric reference value is measured through the flow cell, and a complex reducing agent is added and allowed to react with the available molybdosilicic acid. The reduced molybdosilicic acid has an intense blue color and absorbs light at 810 nm. The intensities of the reference and sample values are compared to a calibration standard, and silica concentration is calculated from

Beer's Law. A similar technique may be applied to *phosphate measurement* at low levels; otherwise standard strip tests are available.

Methods for measuring the total concentration of silica include evaporation and digestion methods and the use of flameless atomic absorption.

4.9. Dissolved Oxygen Electrodes [95-98], [156]

Dissolved oxygen electrodes are an inexpensive, reliable tool of monitoring concentrations of dissolved oxygen in water. The electrodes are two noble metals (for example, silver and gold) immersed in an electrolyte solution. The electrodes are separated from the dectrolyte by a gas-permeable membrane.

The dissolved oxygen permeates across the membrane by partial pressure differences. An electric potential is applied across the electrode and reduces the oxygen inside the membrane. The resulting current is proportional to the oxygen concentration, and the value is calculated by comparing the current to a two-point calibration curve generated from zero oxygen and 100 % humidity air saturation. Sensors vary in quality; typical sensors may accurately measure oxygen concentrations as low as 0.5 µg/L to as high as 7 mg/L.

4.10. Capillary Electrophoresis

Capillary electrophoresis is a variation of electrophoresis (⟶ Electrophoresis – Capillary Zone Electrophoresis) where a sample passes down a capillary. The various components of the sample migrate differentially in an applied electric field and are thus separated. This is a technology with tremendous potential, but has not yet achieved concentration levels required for most ultrapure water application. The technology has been used to analyze anion concentrations to the single-digit µg/L levels. It has been evaluated on site after both power and microelectronics applications, but improved sensitivity is required for wide-spread ultrapure water use.

4.11. ORP Chlorine Analyzers

The oxidation-reduction potential (ORP measured in millivolts) can be used to control free chlorine or free halogen concentrations. There is no direct correlation between the ORP reading and the free chlorine content in water because of other impurities that contribute to the ORP value. One method of control is to use free chlorine test samples to determine the target ORP reading needed to provide the desired free chlorine concentration. The free chlorine concentration should then be tested using a chlorine test method to correlate its value with the ORP controller on a daily basis. ORP is normally measured in the make-up water loop of microelectronics plants.

4.12. Nonvolatile Residue Monitors [120-123]

Nonvolatile residue monitors (NRM) are used to measure dissolved inorganic material, including colloidal silica. The NRM is essentially a very sensitive resistivity monitor capable of measuring resistivity changes induced by charged and uncharged particles.

4.13. Use of On-line Analyzers and Manually Collected Samples

On-line analyzers routinely report lower concentrations than grab samples, because during collection of grab samples materials can be extracted from the sample bottle or atmospheric gases, primarily oxygen and carbon dioxide can be absorbed that interfere with the measurement. However, grab sampling is valuable, not because specific concentrations are observed, but because sampling notes changes in concentrations. The detected changes in baseline concentration measurements indicate problems as they occur.

[Top of Page]

5. Applications

5.1. Microelectronics Industry [167-174], [176-179], [181], [182]

Semiconductor Chip Manufacture (⟶ Semiconductors***).*** Semiconductor chips are manufactured using a complex procedure that chemically imprints a small, rectangular transistor pattern repeatedly on a circular disk, or wafer. After processing, the wafer is cut into rectangular chips, which are placed in a housing for attachment of pin-out connections to the silicon. Although the assembly of the chip requires some rinse water, wafer manufacturing consumes much larger quantities of ultrapure water.

The transistors are applied to the silicon wafer using a combination of film deposition, lithography, ion implantation, and chemical etching in a sequence that may contain more than 100 steps. During many of these steps, chemicals or residuals are rinsed from the wafer. Ultrapure water is used in many processing steps, but the major consumer of ultrapure water is chemical etching, where acids are washed from the wafer.

Since 1985, the number of transistors on processors has doubled nearly every 18 months. The increase in transistors is accomplished, in part, by reducing its line width (or size) from more than 1.0 µm to less than 0.2 µm. As the line width decreases, water purity requirements increase to reduce electrical shorting or physical destruction of the transistors during rinsing.

Semiconductor water systems are complex with many units in series. Figure 5 schematically illustrates the sequence of removal for ionic, organic, particulate, biological, and gaseous contaminants in a semiconductor ultrapure water system.

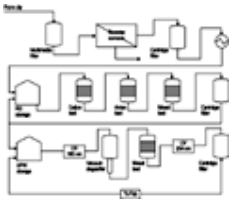

Figure 5. Sample UPW system for semiconductor applications

Fab = fabrication facility

The essential design considerations for semiconductor UPW systems include guaranteed water quality, cost, efficiency, and the prevention of damage to downstream units. The system is composed of three subsystems; the make-up system, the primary loop, and the polish loop.

The Make-Up System. The make-up system removes most contaminants from the inlet water. The primary objective of the make-up system is to maintain the level of the downstream RO storage tank. The primary units used in the make-up loop are:

1. The multimedia filter to remove large particulates, biological contaminants, and organics
2. Reverse osmosis (RO) membranes to remove approximately 99 % of ionic contaminants

A 10 – 20 µm cartridge filter may be used to prevent fouling of the ion exchange resin in the primary loop, and a heat exchanger may be added for temperature control. An additional chemical injection unit (not shown) may be placed before the RO membranes to prevent scaling on the concentrate side of the reverse osmosis units. Special piping materials are not usually required in the make-up system, typical materials of construction are low carbon steel, PVC, or polypropylene. Other purification units may be required in the primary loop if the source water has especially high levels of silica, organics, or biological contaminants.

Primary Loop. The water produced by the primary loop is typically ultrapure resistivity (> 18 MΩ) with respect to ionic contaminants, but organic, particulate, and biological contaminants may still be present. The primary loop serves two purposes: (1) to maintain the level of the downstream ultrapure water tanks, and (2) to dilute the incoming water from the make-up system. The primary loop flow rate is high enough that the water velocity prevents biological growth in the piping.

The main units in the primary loop are:

1. The RO product tank that stores the loop water and accepts make-up water,
2. 185 nm UV lamps, the radiation oxidizes organic contaminants to CO_2 and water.
3. Cation ion exchange resin column that removes metal and other cations,
4. Anion ion exchange resin column that removes halides and other anions,
5. Mixed-bed ion exchange column that removes most other ionic contaminants such as trace amounts of carbonate, silica, borate, and occasionally sodium.
6. 1.0 µm cartridge filter that removes particles that may be generated by the ion exchange resins.

The RO product tank is dimensioned such as to allow ample time to recover from a make-up system malfunction. The ion exchange beds are often chemically regenerated in-situ, i.e. an external resin transfer vessel is not required. The cation exchange bed should always be located upstream of the anion exchange bed. Should the anion exchange bed precede the cation exchange bed, the high pH of the anion effluent may cause metal hydroxide precipitation in the bed, which may foul the resin.

Other configurations of the primary loop are possible, e.g., the cation exchange bed may be removed if the inlet water contains no hardness, or both cation and anion exchange beds may be replaced with a second reverse osmosis stage or an electrodeionization unit. However, most configurations will retain a mixed-bed ion exchange column and a cartridge filter at the end of the primary loop.

Polish Loop. The water quality from the polish loop must meet factory specifications at all times. The key units of the polish loop are:

1. UWP storage tank, blanketed with nitrogen to prevent uptake of atmospheric oxygen and carbon dioxide. Ozone may be introduced in small quantities for biological sterilization and destruction of organics with the 185 nm UV lamps. The tank may be lined with specialty plastics or rubber, and the internals are composed of stainless steel or high quality, low TOC plastics.
2. 185 nm UV lamps, the radiation oxidizes trace organics to carbon dioxide and water. The ozone in the UPW tanks provides the oxygen for the reaction. The primary contaminants leaving the 185 nm UV radiation unit are trace amounts of salts, carbon dioxide, and dissolved gases.
3. Vacuum degasifiers, held at low vacuum to remove the remaining dissolved gases and some of the carbon dioxide. A nitrogen stream is used to vary the oxygen partial pressure in the vacuum and control the concentration of dissolved oxygen.
4. Mixed-bed ion exchange column, the final unit that removes ionic contaminants from the water. Its effluent usually contains only trace amounts of silica and borate. Concentrations of strong electrolytes are usually below detection limits, but the MBIE may release trace amounts of biological and particulate contaminants, especially after years of service.
5. 254 nm UV lamps used for sterilization of biological contaminants leaving the ion exchange columns. The UV

radiation kills almost any microbial contaminant, the dead cells are filtered off.

6. 0.04 µm cartridge filters to remove any particles or biological contaminants that may be generated by the mixed bed, or particles coming from materials of construction that pass through the beds. In 1999, 0.04 µm cartridge filters were the state of the practice. These final filters are positively charged to attract the — primarily — negatively charged particles. As a result, these filters can remove particles less than 0.04 µm in diameter.

The water in the polish loop is circulated continuously, the return flow rate is typically more than two-thirds the supply flow rate. The polish-loop construction materials are stainless steel for the piping from the UPW tank to the vacuum degasifier. After the degasifier, e.g., poly(vinylidene fluoride) (PVDF) piping is required to prevent ionic contamination from metal pipes, TOC leaching from polypropylene or postchlorinated poly(vinyl chloride) (CPVC) and oxygen diffusion from the air typical of other plastic piping. At room temperature, stainless steel pump internals and filter housings are required, while titanium may be used for high temperature applications. Gasket materials are often teflon or silicone.

Performance Issues with Adjacent Equipment. The location of units in the loops is important because removal of one contaminant often leads to generation of another that may be harmful downstream. Other considerations are the effects of a contaminated loop return line, or the accidental introduction of air or chemicals into the system. Some general guidelines are:

1. Reverse osmosis membranes and ion exchange resins may become fouled with particulate or biological matter. A filter should precede a reverse osmosis unit or train of ion exchange columns to prevent fouling.
2. Organic compounds can seriously foul ion exchange resins. Hence, UV lamps should precede the columns so the organic contaminants can be reduced to carbon dioxide. The vacuum degasifier assists in relieving the influx of carbonate to the ion-exchange beds by removing carbon dioxide.
3. Ion exchange resins typically generate particles from cracked resin beads, as well as biological contamination. Filters are needed to remove the particles before the stream is introduced into the loop.

5.2. Pharmaceutical Industry [175]

The U.S.P. defines six types of purified water (http://www.usp.org). These are:

1. Purified water
2. Water for injection
3. Bateriostatic water for injection
4. Sterile water for inhalation
5. Sterile water for injection
6. Sterile water for irrigation

In addition to U.S.P. other agencies or organizations have UPW specifications. Most developed countries have similar organizations that control water quality.

Typically the overall water quality of the water for pharmaceutical applications is less than that required for microelectronic applications, except with regard to microbes and pyrogens. An industrial process may deliver water with a resistivity > 18 MΩ less than 10 µg/L TOC, but no detectable bacteria.

The technologies and materials of construction for these processes are similar to those described for microelectronics UPW. Purification occurs primarily by a combination of RO and DI technologies. Water production rates are much lower than in other applications with total flow through a recirculating loop of a few hundred liters per minute.

5.3. Power Industry [1], [4], [5], [7], [11], [17], [166]

Power plants use ultrapure water to minimize erosion and corrosion of materials within the steam cycle. Water is processed in a make-up plant for all stations. During start-up, additional skid-mounted equipment may be brought to the site for increased capacity. Other applications of water treatment depend on the technology used at the specific plant. Condensate polishing is standard equipment for American nuclear power stations, but many fossil-fired power stations and French nuclear stations do not use condensate polishing within the steam cycle, although it may be required for start-up.

5.3.1. *Nuclear Power Plants*

In addition to make-up water processing, nuclear power plants have several water streams that may require treatment. Plants vary greatly because of different configurations, water chemistry goals, and their location (temperature, seasonal variation, and cooling water source). Some chemicals in the water are more damaging to the materials of construction than others. Consequently, control strategies minimize problem chemicals, such as oxygen and sulfate, at the expense of other chemicals.

The condensate of boiling-water reactors (BWRs) flows directly to the nuclear core (⟶ Nuclear Technology). The steam produced drives the turbines for power generation. The water needs to be very pure, and this generally prevents the addition of chemicals to control corrosion in the circuit. Some BWRs is added to the water to suppress water radiolysis, in others low levels of zinc are added to minimize the build up of cobalt-60 on metal surfaces in the reactor. Primary control strategies within a BWR are to minimize corrosion products, particles and sulfate: both limits are in the single digit µg/L range. To achieve this, full flow filtration and deep bed polishing of concentrates of BWR is employed. Condensate is fed to the steam generator, which maintains a constant-volume liquid level. The boil-up rate results in a concentration factor within the steam generator 100 times the condensate polisher effluent. For example, a 0.3 µg/L sulfate concentration results in a 30 µg/L concentration within the steam generator, which is excessive, and would exceed action limits at most plants and possibly cause corrosion. The reactor water is cleaned using a smaller mixed-bed ion exchange column connected in a recirculation loop with the reactor.

Impurity levels in BWR steam generator water are further controlled by the reactor water clean-up (RWCU) system. Approximately 1 % of the feed flow water is bled from the steam generator and passed through a heat exchanger and then a mixed bed before being returned to the steam generator. The function of the RWCU mixed bed is to remove silica, particulate impurities, sodium chloride, and sulfate. The resins are not regenerated but when exhausted they are transported to radioactive waste storage.

In pressurized water reactors (PWRs) steam generation occurs in a secondary loop, which allows the addition of chemicals, primarily amines, to the water (⟶ Nuclear Technology). The amine addition raises the pH and reduces the corrosion of ferrous metals in the steam – water circuit and hence the transport and deposition of corrosion debris.

Several mixed-bed ion exchange columns are connected in parallel to polish the full-flow of condensate; typically five to seven vessels are required. The control of corrosion debris particles in feed water to a BWR is important since these have high radioactive contamination that may lead to exposure of personnel. Mixed-bed ion exchange columns can function as a particle filter to a limited extent in some plants. However, current technology uses pleated or hollow fiber filters ahead of the mixed beds to remove particles and reduce the loading on the beds.

Beds used as a particle filter must be removed from service frequently for ultrasonic cleaning. This process destroys the concentration loading profile in the exchange column. The result is higher effluent concentrations from the bed with time then desired. Undisturbed beds may remain in service in excess of four years, roughly three to four times longer than beds undergoing regular ultrasonic cleaning.

5.3.2. *Fossil Fuel-Fired Plants*

The principles and methods described for nuclear systems also apply to fossil fuel-fired power plants, but to a lesser degree. Water purification requirements necessary to avoid erosion and corrosion are based on an economic analysis of the entire plant. The economic analysis is clear for the high cost of nuclear facilities, so that investment in water chemistry treatment, or in materials of construction is clearly favorable. However, the initial investment cost of fossil plants is far less. In addition, when maintenance in required, there is no concern for exposure to radioactivity, so maintenance costs are relatively less as well. These factors make condensate polishing less common for fossil-fired plants, although EPRI guidelines still indicate an economic advantage for polishers.

5.4. Chemical Industry

The chemical industry uses ultrapure water primarily for two purposes, boiler feedwater make-up and process water.

5.4.1. *Boiler Feedwater Make-up*

Demineralized water, or an equivalent, is often used for boiler feedwater when the boiler operating pressure is ≥ 6.3 MPa. There are instances when ultrapure water is used as boiler feedwater make-up for boilers operating at 4.24 MPa. Reasons for using ultrapure water for boiler feedwater include:

1. Increasing the cycles of concentration, thus conserving water and energy
2. Reducing the likelihood of scale deposits in the boiler system
3. Systems where the steam is used to drive turbines
4. Systems where the steam is superheated
5. Systems where the steam is injected into a process stream

The acceptable concentration of water impurities decrease as the boiler operating pressure increases. Table 12 describes the maximum allowable conductance, silica concentration, and alkalinity in boiler water as a function of pressure.

Table 12. Maximum silica, alkalinity, and conductance in boiler feedwater as a function of pressure

Pressure, kPa	Total silica concentration, mg/L	Alkalinity, mg/L	Specific conductance, µS/cm
4.240	40	250	2500
5.270	30	200	2000
6.300	20	150	1500
7.000	8	100	1000
10.450	2	none allowed	150
13.900	none allowed	none allowed	100

5.4.2. *Process Water*

There are some applications for ultrapure water in the chemical industry. In most of the cases, specific substances must be kept out of the system to avoid contamination. Uses of ultrapure water in process water applications include the following:

1. Metals rinsing before painting
2. Metals fabrication
3. Process water for metals refining operation
4. Applications where stainless steel will be used at temperatures exceeding 60 °C

5. Process water for pharmaceutical applications

Ultrapure water is often used when one or more of the following common water impurities need to be minimized: calcium, magnesium, sodium, potassium, iron, copper, manganese, chloride, sulfate, silica, alkalinity.

[Top of Page]

6. Future Developments

Ultrapure water quality will continue to be driven by the microelectronics industry. Decreasing line width on semiconductors as computer memory expands and a decrease in processing time dictate these requirements. A common theme of discussion in the industry is that if they can measure contaminants in the water then they should be removed.

Alternatively, chip yield (percentage of working microchips as a fraction of total chips made) should be correlated with water chemistry, and water purity requirements should be established to meet these yield requirements. When taking this approach, added cost to achieve unneeded water quality may be minimized. If sufficient chip yield is achieved with ng/L water quality, then the need for pg/L concentration targets is not justified.

As microelectronic manufacture drives the development of ion exchange and membrane products, all other ultrapure water applications benefit. However, resin and membrane manufacturers must address some specific needs. There is a need for resins that set free a smaller amount of sulfonates, sulfates, and particles. Development of membranes with longer service life will continue, while allowing higher throughput with reduced contaminant permeation. EDI research will continue, and the role of this technology will likely increase. Analytical equipment will be continuously improved with more accurate and lower detection limits. However, the acceptance of this equipment will depend on its cost – benefit ratio.

As the separation technologies increase in performance and asymptotically approach the zero contaminant level, the cost of water production becomes the major design issue; ultrapure water is a utility cost and its minimization leads to greater profits. Waste minimization is also a significant design factor for semiconductor manufacturers. Several UPW plant designs exist using evaporators and crystallizers to attain near-zero-liquid-discharge (NZLD), which allows greater flexibility for site selection because the make-up water requirements are much smaller and environmental impact is significantly decreased.

[Top of Page]

References

General References

1. P. Cohen (ed.): *The ASME Handbook on Water Technology for Thermal Power Systems*, American Society of Mechanical Engineers, New York 1989.
2. W. V. Collentro: *Pharmaceutical Water: System Design, Operation, and Validation*, Interpharm Press, Buffalo Grove IL, 1999.
3. F. G. Helfferich: *Ion Exchange*, McGraw-Hill, New York 1962.
4. M. Henley (ed.): *Ultrapure Water Journal,* Tall Oaks Publishing, Denver, CO, (serial).
5. P. Hoornaert: *Reverse Osmosis*, 1st ed., Pergamon Press, New York 1984.
6. D. Jenkins, V. L. Snoeyink: *Water Chemistry*, John Wiley & Sons, New York 1980.
7. C. J. King: *Separation Processes*, 2nd ed., McGraw-Hill, New York 1980.
8. W. Lorch (ed.): *Handbook of Water Purification*, 2nd ed., Ellis Horwood, Chichester 1987.
9. J. Malievialle, P. E. Odendaal, M. R. Wiesner: *Water Treatment Membrane Processes*, McGraw-Hill, New York 1996.
10. J. A. Marinsky (ed.): *Ion Exchange*, M. Dekker, New York, (serial).
11. T. H. Meltzer: *High Purity Water Preparation*, Tall Oaks Publishing, Denver CO, 1993.
12. D. L. Owens: *Practical Principles of Ion Exchange Water Treatment*, Tall Oaks Publishing, Denver CO, 1985.
13. D. H. Paul: *High Purity Water Treatment*, D. H. Paul, Inc., Farmington NM, 1992.
14. D. Clifford, Z. Zhang: “Modifying ion exchange for combined removal of uranium and radium,” *Journal/American Water Works Association* **86** (1994) 214 – 227.
15. J. Croft, F. Shadman: “Dissolution and removal of colloidal silica in ultra pure water systems,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* **1** (1997) 59 – 70.
16. M. Endo, Y. Nakamura, M. Kogure, T. Fukumoto: “Analysis and removal of newly-found volatile silicon compounds in high-purity water,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1995) 49 – 68.
17. J. Fucsko, M. K. Balazs: “Adsorption and desorption of metals,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1994) 242 – 288.
18. R. Governal, A. Bonner, F. Shadman: “Effect of component interactions on the removal of organic impurities in ultrapure water systems,” *IEEE Trans. Semicond. Manuf.* **4** (1991) 298 – 303.
19. R. R. Harries: “Ion Exchange Kinetics in Condensate Purification,” *Chem. Ind. (London)* (1987) 104 – 109.
20. T. Koloini, P. Rodman, M. Zumer: “Reversible Ion Exchange in Fluidized Beds From Dilute Ternary Solutions,” *Chem. Eng. Commun.* **26** (1984) 297 – 309. Links
21. N. Matsuda, W. Agui, K. Ogino, N. Kawashima, T. Watanabe, H. Sakai, M. Abe: “Disinfection of viable Pseudomonaz stutzeri in ultrapure water with ion exchange resins,” *Colloids Surf. B: Biointerfaces* **7** (1996) 91.
22. N. Matsuo, N. Kawamoto, T. Miyoshi: “Influence of activated carbon microroughness of Si and native oxide on

adsorption of organic carbon in water," *Jpn. J. Appl. Phys., Part 1: Regular Papers & Short Notes & Review Papers* **37** (1998) 4328 – 4329.

23. L. A. Mel'nik, I. A. Grafova, V. D. Grebenyuk, I. I. Penkalo: "Production of high-pure anion exchangers by reagent treatment of technical sorbent AV-17–8," *Khim. Tekhnol. Vody* **13** (1991) 442 – 444.
24. T. Ohmi, T. Isagawa, T. Imaoka, I. Sugiyama: "Ozone decomposition in ultrapure water and continuous ozone sterilization for a semiconductor ultrapure water system," *J. Electrochem. Soc.* **139** (1992) 3336 – 3345. Links
25. D. Peterson, K. Wilson, P. Aicholtz: "Particle control in circulating systems," *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 274 – 290.
26. H. Sawada, T. Hirano, S. Momose, S. Inoue, K. Koba: "High-temperature ultrapure water system based on distillation process," *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1993) 157 – 179.
27. M. J. Slater: "Continuous Ion Exchange: Design Calculations," *Effluent Water Treat. J.* **21** (1981) 461, 464 – 466.
28. A. Tammien, K. Vuorilehto: "Application of a three-dimensional ion-exchange electrolyte in the deoxygenation of low-conductivity water," *J. Appl. Electrochem.* **27** (1997) 1095 – 1099.
29. M. Tanabe, S. Kaneko: "Application of boron selective resin in ultra-pure water system," *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1996) 98 – 107.
30. A. P. Van der Meer, D. N. M. M. Weve, J. A. Wesselingh: *Hydrodynamics and Mass Transfer in a Counter-Current Ion Exchange Pilot Plant Column*, Ellis Horwood Ltd., Chichester 1984, pp. 284 – 294.
31. J. Wibowo, F. Shadman, D. Blackford: "Measuring and removing dissolved and colloidal silica in ultrapure water," *Micro* **15** (1997).
32. J. Wibowo, F. Shadman, D. Blackford: "Measurement and removal of dissolved colloidal silica in ultrapure water," *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* **1** (1997) 43 – 58.
33. Y. Yagi, F. Hayashi, Y. Uchitomi: "Evaluation of boron behavior in ultrapure water manufacturing system," *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1994) 54 – 63.
34. Y. Yagi, K. Takino, Y. Uchitomi: "Study on the sterilization of adhering bacteria in ultrapure water manufacturing system," *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1993) 63 – 74.
35. M. Ball, R. R. Harries: "Resins for high-purity water production," *J. Chem. Technol. Biotechnol.* **45** (1989) 97 – 107. Links
36. J. R. Emmett: "Anion Resin Kinetics At High Flow Rates," *NATO ASI Series, Series E: Applied Sciences* (1985) 41 – 55.
37. R. Fernandez-Prini, C. R. de Pattin, P. Schulman: "Hydrothermal Decomposition of Ion-Exchange Resins," *Power Industry Research* **2** (1982) 101 – 108.
38. G.-C. Lee, G. L. Foutch, A. Arunachalam: "Evaluation of mass-transfer coefficients for new and used ion-exchange resins," *Reactive and Functional Polymers* **35** (1997) 55 – 73. Links
39. J. Loureiro, C. Costa, M. Dias, J. Lopes, A. Rodrigues: "Design Methods For Ion Exchange Equipment," *NATO ASI Series, Series E: Applied Sciences* (1985) 245 – 260.
40. Z. Matejka, V. Vlckova, M. Jileckova: "Ion-Exchange Kinetics of Strong-Base Anion Exchangers in Condensate Treatment," *React. Polymers, Ion Exchangers, Sorbents* **7** (1986) 319.
41. B. Wang, Z. Ding, B. He: "Characteristics of ion exchange resin with uniform particle size," *Ion Exchange Absorption/Lizi Jiaohuan Yu Xifu* **9** (1993) 168 – 173.
42. J. C. Bates, T. D. Johnson: *Development of a Computer Model, Amm Leak, for Sodium Leakage from* NH_4^+ OH^- *form Mixed Beds*, Ellis Horwood Ltd., Chichester 1984, pp. 213 – 241.
43. V. Chowdiah, G. L. Foutch: "Kinetic model for cationic-exchange-resin regeneration," *Ind. Eng. Chem. Res.* **34** (1995) 4040. Links
44. F. X. McGarvey: "Application of ion exchange to meet the many challenges to produce an improved grade of pure water in the 90's," *ASTM Spec. Tech. Pub.* (1991) 334 – 340.
45. D. Petruzzelli, G. Tiravanti, R. Passino: "Cr(III)/Al(III)/Fe(III) ion binding on mixed bed ion exchangers. Synergistic effects of the resins behaviour," *Reactive and Functional Polymers* **31** (1996) 179 – 185. Links
46. H. Takase, F. Fukuoka: "Estimation Method For Exchange Capacity of Mixed Bed Using Monitoring Minicolumns," *J. Nuclear Sci. Technol.* **23** (1986) 338 – 351.
47. S. V. Divekar, G. L. Foutch, C. E. Haub: "Mixed-Bed Ion Exchange At Concentrations Approaching the Dissociation of Water. Temperature Effects," *Ind. Eng. Chem. Res.* **26** (1987) 1906 – 1909. Links
48. G. L. Foutch, R. Bulusu, D. J. Morgan, "The Fate of Bicarbonate/Carbonate in Mixed-Bed Ion Exchange," *Ultrapure Water* **12** (1995) 58.
49. G. L. Foutch, B. I. Noh, V. Chowdiah: "Integration of mathematical models into on-line operation of mixed-bed ion exchange," *Proc. Frontiers of Power Conference* (1991) XIV-1 – XIV-5.
50. E. J. Zecchini, G. L. Foutch: "Mixed-bed ion-exchange modeling with amine form cation resins," *Ind. Eng. Chem. Res.* **30** (1991) 1886 – 1892. Links
51. J. A. Mathews: "Evaluation of Weak Electrolyte Resins For Applications in a Mixed Resin System," *Proc. Int. Water. Conf. Eng. Soc. West. P.* (1985) 44 – 52.
52. S. Sourirajan, T. Matsura: *Encyclopedia of Fluid Mechanics*, Gulf Publishing Co., Houston TX, 1986.
53. R. Singh: "Membranes: A Review of Membrane Technologies: Reverse Osmosis, Nanofiltration, and Ultrafiltration," *Ultrapure Water* **14** (1997) 21 – 29.
54. D. Baker: "Ion exchange and membrane processes," *Process Control Eng.* **45** (1992) 40 – 44.
55. D. Pines, P. Birbara: "Ultrapure water processing system utilizing membrane pervaporation and catalytic oxidation technologies," *SAE Trans.* **100** (1991) 1926 – 1934.

56. M. Xu, W.-C. Li, H.-C. Wang, H. Rahman: “Particle shedding characteristics of membrane filters under pulsed-flow conditions in liquid filtration,” *Inst. Environ. Sci. Proc. Annu. Tech. Meet.* (1998) 178 – 183.
57. W. Byrne: *Reverse Osmosis: A Practical Guide for Industrial Users*, Tall Oaks Publishing Inc., Littleton, CO 1995.
58. G. Pittner: *Reverse Osmosis: Membrane Technology*, Water Chemistry, and Industrial Applications, Van Nostrand Reinhold, New York 1993.
59. Z. H. Amjad, J.J. Pugh: “Reverse Osmosis Failures: Causes, Cases, and Prevention,” *Ultrapure Water* **8** (1991) 14 – 21.
60. M. Bukay: “How to Monitor a Reverse Osmosis System,” *Ultrapure Water* **1** (1984) 32 – 36.
61. J. C. Carmody, A. R. Lindahl, J. E. Martyak: “Meeting the challenges of RO/DI system contamination control in the 1990s,” *Microcontamination* **8** (1990) 29 – 32, 62.
62. T. Ikeda, M. Kurihara: “High Performance Composite Reverse Osmosis Membranes: Formation of Composite Membranes and Their Application for Ultrapure Water Manufacturing,” *Denshi Joho Tsushin Gakkai Shi/J. Inst. Electron. Inf. Commun. Eng.* **77** (1994) 1139.
63. T. Ikeda, H. Muragishi, R. Bairinji, T. Uemura: “Advanced reverse osmosis membrane modules for novel ultrapure water production process,” *Desalination* **98** (1994) 391 – 400. Links
64. M. Okazaki, T. Ikeda, Y. Nakagawa, R. Bairinji: “Recent progress of revers osmosis membrane modules for ultrapure water production,” *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 87 – 95.
65. Arrowhead Industrial Water Inc., US 4 574 049, 1986.
66. Zenon Environmental Inc., US 5 766 479, 1998.
67. SEH America, Inc., US 5 651 894, 1997.
68. A.-S. Jonsson, G. Tragardh: “Ultrafiltration applications,” *Desalination* **77** (1990) 135 – 179.
69. Matsuo: “Novel ultrafiltration module as a final filter in the ultrapure water system,” *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 108 – 127.
70. Illinois Water Treatment, Inc., US 5 518 624, 1996.
71. F. DiMascio, J. Wood, J. M. Fenton: “Continuous electrodeionization: production of high-purity water without regeneration chemicals,” *Electrochem. Soc. Interface* **7** (1998) 26 – 29.
72. G. C. Ganzi, A. D. Jha, F. DiMascio, J. H. Wood: “Theory and Practice of Continuous Electrodeionization,” *Ultrapure Water* **10** (1997) 64 – 69.
73. G. C. Ganzi, J. H. Wood, C. S. Griffin: “Water purification and recycling using the CDI process,” *Environ. Progr.* **11** (1992) 49 – 53.
74. Christ AG, US 5 376 253, 1994.
75. Millipore Corporation, US 4 925 541, 1990.
76. Ebara Corporation, Atomic Energy Research Institute, US 5 308 467, 1994.
77. IP Holding Company, US 5 316 637, 1994.
78. Millipore Corporation, US 4 753 681, 1988.
79. Millipore Corporation, US 5 154 809, 1992.
80. V. L. Snoeyink, M. T. Suidan: *Disinfection: Water and Wastewater*, Ann Arbor Science, Ann Arbor, MI 1975.
81. T. H. Meltzer: “Maximization of filtrative particle removals,” *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 165 – 175.
82. R. Hango: “Performance testing submicron filters for electronics grade ultrapure water,” Proceedings of the Membrane Conference and High Tech Separations Symposium, Business Communications Co., Nov. 4 – 6, 1991, 369.
83. T. Hori, M. Hashino, A. Omori, T. Matsuda, K. Takasa, K. Watanabe: “Synthesis of novel microfilters with ion-exchange capacity and its application to ultrapure water production systems,” *J. Membr. Sci.* **132** (1997) 203 – 211.
84. H. G. Schroeder: “Filter performance evaluation. A comprehensive approach,” *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 176 – 193.
85. Spinnstoffabrik Zehlendorf AG, US 4 077 880, 1978.
86. The Ducon Company, Inc., US 3 798 882, 1974.
87. US 3 912 466, 1975 (A. F. Zenz).
88. A. Shucosky: “High flux membrane cartridge filter,” *Filtr. Sep.* **29** (1992) 23 – 27.
89. T. Saitoh, H. Nagai, M. Imai, K. Nakajima, M. Tujimura: “Simultaneous removal of DO and TOC in ultrapure water by using UV rays,” *Desalination* **98** (1994) 249 – 255. Links
90. G. Chen, F. Shadman: “Removal of trace surfactant from ultrapure water though UV pyrolysis and photocatalytic oxidation,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1996) 111 – 124.
91. H. Kano, K. Oda, M. Aoyagi: “Behavior of ‘colloidal substances’ in ozone-UV treatment,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1994) 21 – 39.
92. Kawada, J. Tanaka, N. Uchiyama, H. Yagi, Y. Toriyama, A. Umeka: “Non-ionic silica removal using ozone-UV treatment for semiconductor ultrapure water systems,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1994) 1 – 20.
93. B. Srikanth, M. S. Chem: “Recent advancements in UV technology yield enhanced TOC reduction performance,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1997) 111 – 130.
94. *Nalco Water Handbook*, Nalco Chemical Co., Naperville, IL 1979, pp. 17–6.

95. Bujedo, P. A. Peterson: “Removing dissolved oxygen from ultrapure water,” *Semicond. Int.* **20** (1997) 137 – 138.
96. Y. Kasama, Y. Yagi, T. Imaoka, M. Kawakami, T. Ohmi: “Advanced D.I. water system with low dissolved oxygen for ULSI processing,” *Proc. Annu. Tech. Meet. Inst. Environ. Sci.* (1990) 344 – 349.
97. Y. Kasama, Y. Yagi, T. Imaoka, T. Ohmi: “Low dissolved oxygen ultrapure water systems for native oxide free wafer processing,” *Conf. Solid State Devices Mater.* (1990) 1139 – 1142.
98. Li, I. Chua, W. J. Ng, W. K. Teo: “Removal of dissolved oxygen in ultrapure water production using a membrane reactor,” *Chem. Eng. Sci.* **50** (1995) 3547 – 3556. Links
99. H. Miyamaru, T. Takeda, K. Oda, M. Aoyagi: “Influence of dissolved oxygen on the total organic carbon in ultrapure water,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1993) 75 – 92.
100. H. Sato, N. Hashimoto, T. Shinoda, K. Takino: “Dissolved oxygen removal in ultrapure water for semiconductor manufacturing,” *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 147 – 164.
101. S. L. Tai, I. Chua, K. Li, W. J. Ng, W. K. Teo: “Removal of dissolved oxygen in ultrapure water production using microporous membrane modules,” *J. Mem. Sci.* **87** (1994) 99 – 105.
102. Y. Yagi, T. Imaoka, Y. Kasama, T. Ohmi: “Advanced ultrapure water systems with low dissolved oxygen for native oxide free wafer processing,” *IEEE Trans. Semicond. Manuf.* **5** (1992) 121 – 127.
103. A. Sengupta, P. A. Peterson, B. D. Miller, J. Schneider, C. W. Fulk, Jr., “Large-scale application of membrane contactors for gas transfer from or to ultrapure water,” *Sep. Purif. Technol.* **14** (1998) 189 – 200. Links
104. Hoechst Celanese Corp., US 5 695 545, 1997.
105. Rhône-Poulenc, US 3 651 616, 1972.
106. Y. Ishihara, S. Yamane, H. Yamazaki, H. Tsuge: “Development of nitrogen gas bubbling deaerator made of polyvinyliden fluoride,” *J. Electrochem. Soc.* **142** (1995) 2352 – 2357. Links
107. Burkhart, J. Wermelinger, F. Klaiber: “Is polyvinylidene fluoride piping safe for hot ultrapure-water applications?” *Microcontamination* **13** (1995) 27 – 31.
108. M. Burkhart, J. S. Wermelinger, W. D. Muller: “Suitability to Polyvinylidiene Fluoride (PVDF) Piping in Pharmaceutical Ultrapure Water Applications,” *PDA J. Pharmaceut. Sci. Technol.* **50** (1996) 246 – 251.
109. Y. Kajiyama, M. Sugisawa, H. Chifuku, K. Ushikoshi: “Piping technologie for passivated stainless steel,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1993) 196 – 215.
110. A. Miyake, T. Isagawa: “Release behavior of metal ions from passivated stainless steel in hot ultrapure water,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1995) 23 – 40.
111. Y. Motomura, K. Yabe: “Piping materials and distribution systems for advanced ultrapure water,” *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 1 – 22.
112. D. Sinha: “Surface study of polyvinylidene fluoride pipe exposed to high-purity water containing dissolved ozone,” *J. Electrochem. Soc.* **144** (1997) 2840 – 2843. Links
113. K. Ushikoshi *et al.*: “Ultrapure water treatment system made of passivated stainless steel (GOLD EP),” *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 23 – 41.
114. S. E. Carpenter: “Improving epifluorescent monitoring methods in estimating numbers of bacteria in ultrapure water,” *Proc. Annu. Meet. Inst. Environ. Sci.* (1990) 212 – 216.
115. T. Chu, M. K. Balazs: “Determination of total silica at PPB levels in ultrapure water by three different analytical techniques,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1993) 140 – 156.
116. L. Fabry, S. Pahlke, L. Kotz, P. Bloechel, A. Kempf, T. Ehmann, K. Baechmann: “Creative engineering and ultrapure-trace analysis of ultra pure process media,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1996) 1 – 24.
117. G. C. Frye, D. S. Blair, T. W. Schneider, C. D. Mowry, C. W. Colburn, R. P. Donovan: “Development and evaluation of on-line detection techniques for polar organics in ultrapure water,” *Proc. Annu. Tech. Meet. Inst. Environ. Sci.* (1996) 355 – 362.
118. T. Imaoka, A. Umeka, T. Futatsuki: “Identification of trace organic impurities in ultrapure water and influence on silicon wafer surface,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1993) 93 – 116.
119. K. Kitami, T. Mizuniwa, M. Ito, R. Miwa, K. Yabe: “Ultrapure water quality and analytical technology for advanced semiconductor manufacturing,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1997) 71 – 96.
120. T. Nogami, T. Ohto, O. Kawaguchi, Y. Zaitsu, S. Sasaki: “Estimation of bacterial contamination in ultrapure water: Application of the anti-DNA antibody,” *Anal. Chem.* **70** (1998) 5296 – 5301. Links
121. I. L. Pepper *et al.*: “Rapid and systematic analytical method for measuring bacterial contaminats in ultrapure water,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1993) 50 – 62.
122. I. L. Pepper *et al.*: “Measuring bacterial contaminats in ultrapure water. A rapid analytical method,” *Microcontamination* **12** (1994).
123. P. Rodacy, P. Leslie, S. Klassen, R. Silva: “Trace identification of organic molecules in ultrapure water using ion mobility spectroscopy,” *Annu. Tech. Meet. Inst. Environ. Sci. Proc.* (1996) 363 – 367.
124. M. A. Sadler, R. Bolton, E. K. Bullas: “PPQ water-the next challenge,” *Proc. Int. Water Conf. Eng. Soc. West. P.* (1992) 16 – 25.
125. J. Schmitt, H.-C. Flemming: “FTIR-spectroscopy in microbial and materials analysis,” *Int. Biodeterior. Biodegrad.* **41** (1998) 1 – 11.
126. J. Weiss: *Handbook of Ion Chromatography*, VCH Verlagsgesellschaft, Weinheim, Germany 1995.
127. B. A. Kyereboah-Taylor: “Ultrapure water for ion chromatography,” *Am. Lab.* **27** (1995) 24A. Links

128. T. Taniwaki, J. Nakano, T. Kaneko, Y. Itoh: “Ultra-trace analysis of anions for ultrapure water using ion chromatography,” *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1993) 216 – 230.
129. C. A. Pohl, E. L. Johnson: “Ion Chromatography — State of the Art,” *J. Chromatogr. Sci.* **18** (1980) 442 – 452. Links
130. Dow Chemical, US 3 923 460, 1975.
131. Dow Chemical, US 3 920 397, 1975.
132. Dionex Corporation, US 5 597 734, 1997.
133. Dow Chemical, US 4 751 004, 1988.
134. US 5 376 254, 1994 (A. V. Fisher).
135. SenTech Medical Corp., US 4 549 951, 1985.
136. Hitachi, Ltd., US 5 112 471, 1992.
137. M. J. Bollinger, C. A. Craig, R. Godec, P. Kosenka, K. O'Neil: “Novel approach to verifying TOC instrument accuracy,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1997) no. 1, 97 – 110.
138. R. Governal, A. Bonner, F. Shadman: “Effect of system interactions on the removal of TOC from DI water polishing loops,” *Proc. IEEE/CHMT 9th Int. Electronic Manuf. Technol. Symp.* (1990) 141 – 144.
139. H. Morita, Y. Obata, M. Koizumi: “Advanced oxidation process for TOC removal in ultra pure water system,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1996) 53 – 67.
140. R. Wen: “Manufacture of ultrapure water with low concentrations of TOC and bacterial endotoxins,” *Chinese J. Electron.* **7** (1998) 158 – 161. Links
141. C. D. Jolly, E. L. Jeffers: “Ultrapure water total organic carbon analyzer — advanced component development,” *SAE Tech. Paper Ser.* (1991) 131 – 137.
142. M. Nitta, T. Iwata, Y. Sanui, T. Ogawachi: “Determination of total organic carbon in highly purified water by wet oxidation at high temperature and high pressure,” *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 314 – 320.
143. Anatel Corp., US 5 395 522, 1995.
144. Anatel Instrument Corp., US 4 666 860, 1987.
145. US 5 413 763, 1995 (J. Jeffers).
146. US 5 518 608, 1996 (H. Chubachi).
147. H. Fujimori, T. Matsui, T. Ajiro, K. Yokose, Y.-M. Hsueh, S. Izumi: “Detection of fine particles in liquids by laser breakdown method,” *Jpn. J. Appl. Phys., Part 1: Regular Papers & Short Notes* **31** (1992) 1514 – 1518.
148. J. Livingston: “Use of a real-time liquid particle counter as a monitor of ultrapure water quality at point of use,” *Proc. Annu. Tech. Meet. Inst. Environ. Sci.* (1990) 308 – 311.
149. T. Mizuniwa, K. Yabe: “0.05-micron particle counting using scanning electron microscope,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1993) 37 – 49.
150. Hitachi, Ltd., US 5 178 836, 1993.
151. BASF AG, US 4 752 131, 1988.
152. Shinmikuni Kikai Ltd., US 5 534 999, 1996.
153. Hydac Technology GmbH, US 5 122 655, 1992.
154. C. L. Hillard: “Instruments: Colormetric and Spectrophotometric Analysis for Water,” *Ultrapure Water* **14** (1997) 25 – 29.
155. US 5 098 547, 1992 (A. I. Bryan, M. R. Cushman).
156. Orbisphere Laboratories Neuchatel, US 5 144 831, 1992.
157. Scottish & Newcastle Breweries, Ltd., US 4 461 165, 1984.
158. Hitachi, Ltd., US 5 730 850, 1998.
159. Princeton Biochemicals, Inc., US 5 045 172, 1991.
160. D. B. Blackford, T. A. Kerrick: “Nonvolatile residue measurement of small volume solvents,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1994) 76 – 83.
161. T. Mizuta, I. Sugiyama, T. Nakai, Y. Ohta: “Quality control of ultrapure water using high sensitivity total residue monitor,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1993) 231 – 256.
162. TSI Inc., US 5 351 523, 1994.
163. G. Bellotti: “Future Criteria For the Design and Manufacturing of Water Treatment Plants,” *NATO ASI Ser., Series E: Applied Sciences* (1985) 261 – 272.
164. T. Futatsuki, N. Urai, I. Shindo: “Wastewater reclamation systems and closed water systems,” *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 42 – 68.
165. F. M. Cutler, D. C. Auerswald, S. Simmons: “Parts per trillion with separate bed polishing,” *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 194 – 226.
166. J. G. Grantham: “Resins For Condensate Polishing,” *Effluent Water Treat. J.* **25** (1985) 195 – 198.
167. R. Hanselka: “Zero unscheduled downtime for modern ultra pure water systems,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1995) 41 – 48.
168. T. Isagawa, M. Kogure, T. Imaoka, T. Ohmi: “Ozone added ultrapure water applications for ULSI advanced processing,” *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1992) 224 – 249.
169. J. Jenkins, B. Pennington: “Tracking, identifying, and resolving intermittent particle excursions in a large UPW

system," *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1996) 49 – 70.

170. T. Lockwood, G. Waddington, P. Parise: "No chemical waste ultrapure water system," *Proc. Int. Water Conf. Eng. Soc. West. P.* (1990) 50.
171. N. Matsuda, W. Agui, T. Tougou, H. Sakai, K. Ogino, M. Abe: "Gram-negative bacteria viable in ultrapure water: identification of bacteria isolated from ultrapure water and effect of temperature on their behavior," *Colloids Surf. B: Biointerfaces* **5** (1996) 279 – 289.
172. T. Mizuniwa, K. Yabe: "Improvement of ultrapure water quality. Reduction of particles," *Proc. Electrochem. Soc.* **90** (1990) 173 – 180.
173. R. Mohindra, W. Kern: "New process for producing particle-free deionized water," *Semicond. Int.* **20** (1997) 191 – 192, 194, 196, 198.
174. H. Morita, J. Ida, T. Mizuniwa, T. Ohmi: "Hydrogenated ultrapure water production system for future wet cleaning process," *Diffusion Defect Data B: Solid State Phenomena* **65 – 66** (1999) 7 – 10.
175. J. F. Nykanen, R. M. Cutler: "Designing and operating a pharmaceutical ultrapure water system," *Microcontamination* **8** (1990).
176. T. Ohmi: "Ultrapure water production/supply technology in support of ULSI manufacturing," *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 251 – 273.
177. L. Peters: "Ultrapure water: Rewards of recycling," *Semicond. Int.* **21** (1998) 71 – 72, 74, 76.
178. G. A. Pitts: "Construction issues facing ultra pure water systems," *ASCE Construct. Congr. Proc.* (1997) 202 – 208.
179. Y. Shiramizu, H. Kitajima: "Requirements of ultra pure water and chemicals for wet cleaning," *Conference Proceedings — Annual Semiconductor Pure Water and Chemicals Conference* (1995) 80 – 100.
180. H. Soyama: "Cavitation characteristics of ultrapure water (Behavior of cavitation nuclei in vibratory erosion test)," *Nippon Kikai Gakkai Ronbunshu, B Hen/Trans. Jpn. Soc. Mechan. Eng. B* **60** (1994) 6 – 11.
181. B. M. Stewart, D. Darbouret: "Advancements in the production of ultrapure water for ICP-MS metals analysis," *Am. Lab.* **30** (1998) 36 – 38. Links
182. Y. Yamaki, Y. Ohta: "New concepts in on-line instrumentation," *Conference Proceedings — Annual Semiconductor Pure Water Conference* (1991) 275 – 313.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Bioremediation

(see also → Soil, Chap. 7.1.2.2.3.)

GUSTAV ADOLPH HENKE, Umweltschutz Nord GmbH & Co, Ganderkesee, Germany

1. Introduction 1
2. Microbiology 1
3. Fundamentals of Biodegradation Processes 3
4. Processes 5
4.1. On-Site Methods 5
4.2. In-Situ Methods 6
4.3. Off-Site Methods 7
5. Use of the Cleaned Soil 7
6. Case Examples 7
6.1. Remediation of a Tank Farm 7
6.2. Remediation of Soil Contaminated with TNT 8
7. References 10

1. Introduction

Bioremediation designates soil remediation methods, which make use of the natural metabolic potential of microorganisms to degrade organic contaminants. In order to achieve maximum degradation rates optimal environmental conditions (provision of nutrients, moisture, temperature) must be set. The contaminated soil to be cleaned is either excavated or, if this is not possible, treated by in-situ methods.

The development of biological soil remediation, from initial tests on a small scale to cleaning large quantities of soil, took place over a period of about 15 years. Three large phases can be distinguished:

Phase 1. At the beginning exclusively static treatment of the contaminated soil was carried out. After technical preparation of the soil (sorting, crushing, addition of nutrients, homogenization) open bio-beds exposed to the weather were laid out (Fig. 1). The remediation periods for degradation of hydrocarbons were approximately 1 – 2 years.

Phase 2. In the next step the bio-beds were provided with covers to become independent of the weather (Fig. 2). Furthermore, shorter degradation periods could be achieved during the warm season due to the increased temperature under the cover. In addition, the remediation process was accelerated by regular turning of the contaminated soil.

Phase 3. On-site biological soil remediation was further optimized by using large tents instead of low covers. In this way, large quantities of soil could be cleaned and larger turning machines could be employed. Further development led to the building of stationary soil remediation plants in which more than over 100 000 t of contaminated soil are being cleaned today. In 1999, the throughput of biological soil remediation plants in Germany was around 2.7 times higher than that of soil washing plants [4].

The scope of application of the virtually natural remediation technology now encompasses a wide range of contaminants. In addition to the "customary" damage caused by petroleum contamination, polycyclic aromatic hydrocarbons (PAH) [5–7], highly volatile chlorinated hydrocarbons (CHC) [8], explosives such as trinitrotoluol (TNT) [9] and other organic contaminants can be detoxified by biological treatment.

2. Microbiology

Microorganisms are usually unicellular organisms with small dimensions and a simple cell structure. Microorganisms include algae, fungi and protozoa, which can be classified as *eukaryotes*, as well as *bacteria* in a restricted sense and *archaebacteria* (*prokaryotes*), whose cell structure differs considerably from that of all other organisms. For biological remediation methods the decomposition induced by bacteria and fungi in the soil is of primary importance.

Because of their special genetic structure *bacteria* are able to change their genetic charac-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.d04_d01

Figure 1. Biological soil remediation with open bio-bed

Figure 2. Biological soil remediation with closed bio-bed

ter abruptly, i.e., to mutate. Their spontaneous "shifts in genetic characteristics" and very short reproduction cycles impart these microorganisms a significant adaptability to their surroundings. In this way "nutrients" that previously could not be utilized may also be degraded microbially after a while.

Fungi occur in a wide variety of forms, ranging from microscopically small, unicellular organisms to visible fungi → Fungicides, Agricultural. Since fungi do not have photosynthetic pigments as green plants, they have to feed on organic material. Most fungi grow under aerobic conditions and obtain energy by decomposition of organic substances.

Wood-degrading "white rot" fungi, which belong to the group of basidiomycetes, are of special importance for biological remediation methods. These fungi, such as *Trametes versicolor* (Fig. 3), *Pleurotes ostreatus* and *Phanerochaete*

chrysosporium, are able to cleave the chemical structure of lignin with the help of special enzymes.

Figure 3. The "white rot" fungus *Trametes versicolor* [10]

Tar oils, i.e., PAH, which frequently occur as soil contaminants in the ground of gas works, coking plants, and companies that process tar products, have a similar chemical structure to lignin. These substances are often difficult to decompose by bacteria due to their high chemical stability and low solubility in water. White rot fungi have been successfully used several times for biodegradation of PAH with lower molecular mass thanks to the similarity in the chemical structure of the substrates [11–13]. However, high-molecular PAHs are more or less nondegradable. Furthermore, the successful biological decomposition of tar oil is highly dependent on the soil structure and the concentration of organic carbon.

3. Fundamentals of Biodegradation Processes

Many microorganisms are able to decompose not only biologically generated substances, e.g., petroleum hydrocarbons, but also synthetic organic substances. By virtue of the high degree of flexibility with respect to metabolism and the adaptability of a large number of bacteria and fungi, the scope of microbial degradation theoretically encompasses all groups of organic contaminant groups. An analysis of examined contaminated sites showed that most types of contamination are of organic nature (Fig. 4) and can thus be biologically degraded in theory.

The organic substance can be decomposed completely (mineralized) or transformed and thus be used to build up microbial biomass. Furthermore, the organic molecules can be recycled to the natural carbon store in the soil (humic matter) depending on the degree of their degradability. Delayed mineralization of the organic substance then takes place in the very slow process of humic metabolism.

The possibility of microbial degradation has been verified for a great number of environmental contaminants by means of laboratory and field tests. For many aliphatic and aromatic mineral oil hydrocarbons the degradation by bacteria and the regulation mechanisms of this degradation have already been elucidated completely.

One of the most important condition for the successful degradation of aromatic hydrocarbons is the availability of oxygen, which must be introduced at two steps.

1) The pathway from, e.g., benzene to pyrocatechol requires oxygen to "produce" this more water-soluble substance due to the introduction of two new OH groups (see Fig. 5)
2) The aromatic ring is cleaved only in the presence of oxygen

After the cleavage and the formation of intermediate products the tricarboxylic acid cycle splits the metabolites into low molecular organic acids such as acetic or succinic acid. The final products are carbon dioxide and water. (Fig. 5).

The ability to degrade hydrocarbons is widespread among bacteria and yeasts, but occurs particularly frequently among representatives of the genera *Pseudomonas, Alcaligenes* and *Acinetobacter*. Important HC-degrading bacteria are listed below [14]:

Acinetobacter
Pseudomonas
Leucothrix
Alcaligenes
Spirillium
Micromonospora
Bacillus
Thermomicrobium
Mycococcus
Corynebacterium
Xanthomonas
Klebsiella
Actinomycetes
Sphaerotilus
Micrococcus

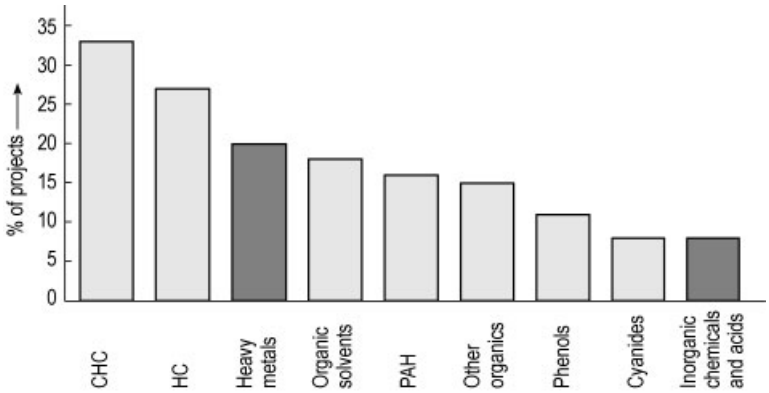

Figure 4. Spectrum of contaminants at contaminated sites
CHC = chlorinated hydrocarbons; HC = hydrocarbons; PAH = polycyclic aromatic hydrocarbons

Arthrobacter
Streptomyces
Mycobacterium
Brevibacterium
Vibrio
Nocardia

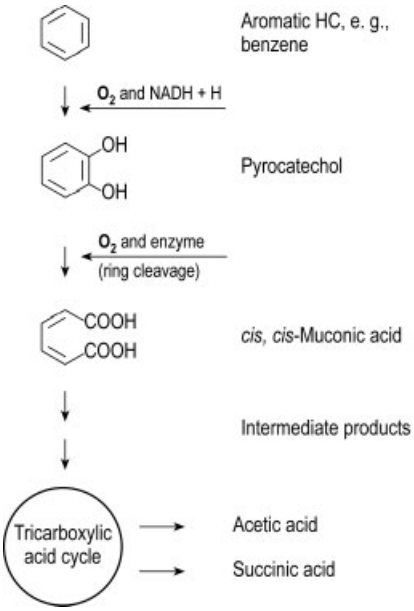

Figure 5. Schematic diagram of degradation of aromatic hydrocarbons

Important HC-degrading fungi and yeasts include [14]:

Aspergillus
Phoma
Gliocladium
Beauveria
Scopulariopsis
Hansenula
Citromyces
Rhodotorula
Mortierella
Debaryomyces
Torula
Penicillium
Fusarium
Aureobasidium
Scolecobasidium
Graphium
Candida
Sporobolomyces
Monilia
Chladosporidium
Tolypocladium
Paecilomyces
Endomyces
Torolopsis

These phyla could be isolated from a wide variety of ecosystems (seawater and fresh water, soil, sediment) [15]. The microorganisms additionally showed extremely high degradation rates for polycyclic cycloalkanes and aromatics. In ecosystems that have not yet been in contact with petroleum contaminations, petroleum petroleum-decomposing phyla represent only a small proportion of the microbial population whereas they can make up to 100 % of the microorganisms in contaminated areas. The percentage of organisms that degrade petroleum hydrocarbons can thus be regarded as a sensitive

biological indicator for hydrocarbon contaminations [15].

Some difficultly degradable compounds – high molecular condensed aromatics, highly chlorinated organics, dioxins, dibenzofurans – can be partially microbially degraded, predominantly under special laboratory conditions [6]. These tests have shown that the limits of biological decontamination processes are not determined by the metabolic abilities of the bacteria, but to a great degree by the environmental conditions (abiotic factors).

Generally, microbiological remediation methods make use of the natural metabolic potential of microorganisms. The soil can be viewed as a biological fixed-bed reactor with a large internal surface. Microbial metabolic activities are fostered by optimal provision of mineral nutrients, moisture, and temperature conditions.

The supply of microorganisms with oxygen is ensured by introduction of hydrogen peroxide and ozone or by mechanically working the soil (turning, ploughing, tilling). In addition, mineral nutrients (especially phosphorus and nitrogen) have to be supplied. Nitrogen and phosphorus must be added in the correct relation to one another and to the concentration of organic carbon in order to provide for extensive mineralization of the contaminants. An optimal pH value (for bacteria in a slightly alkaline range of 7.8 – 8) can be achieved by addition of lime. Another prerequisite for a high degradation rate is a uniform distribution of contaminants along with good permeability of the soil, which is achieved in remediation measures through preparation of the excavated soil (sorting, homogenization).

4. Processes

Treatment of contaminated soil is either carried out locally (on-site) or in soil remediation plants (off-site). If the contaminated soil cannot be excavated, in-situ methods are applied (see Fig. 6).

4.1. On-Site Methods

Bio-Beds. Biological soil remediation can be carried out on-site primarily in regeneration heaps/bio-beds/biopiles or in bioreactors. The schematic process is presented in Figure 7. After excavation of the soil the material is homogenized to ensure:

Uniform distribution of contamination
Loose soil structure
Good aeration of the soil, and
Optimal availability of contaminants for the microorganisms

During homogenization various additives, such as bark mulch, straw or compost material are added, according to the need for structural improvement.

After pretreatment the actual microbiological degradation takes place. It is of crucial importance that the microorganisms working and reproducing in the soil are supplied with sufficient oxygen. Mainly biodegradation is started with a *dry rotting* process set up without base seal. In the dry rotting process air is either introduced through regular turning of the soil or through aeration pipes running inside the beds. In the sprinkler method the bio-beds are equipped with a sprinkler device and appropriate sealing at the bottom. In the case of the sprinkler method the oxygen reaches the bacteria via the water.

Environmental conditions (e.g. water content, temperature, pH value) are regularly monitored in a laboratory.

Bioreactors. In addition to the heap methods, treatment in closed bioreactors is being increasingly implemented for degradation of more complex contaminants. Under completely controllable conditions, the following objectives are achieved in closed bioreactors:

Acceleration of contaminant degradation
Better bioavailability of the contaminants
Better contact between contaminants and bacteria, and
Application of biodegradation to new substance classes

In bioreactors chlorinated hydrocarbons [16] or TNT can be detoxified by biodegradation, transformation, or through an alternation between anaerobic and aerobic conditions.

Since bioreactors are very complex and costly devices, their use only makes sense if no other methods are available.

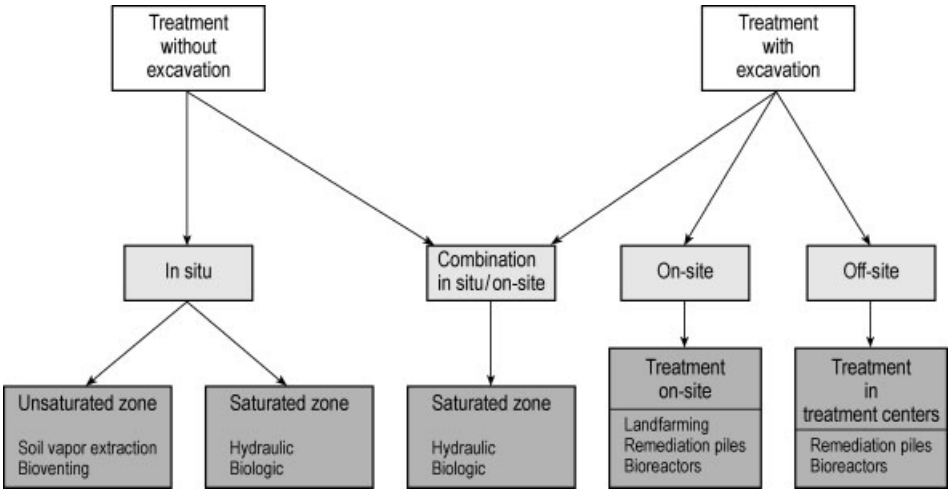

Figure 6. Microbiological remediation methods

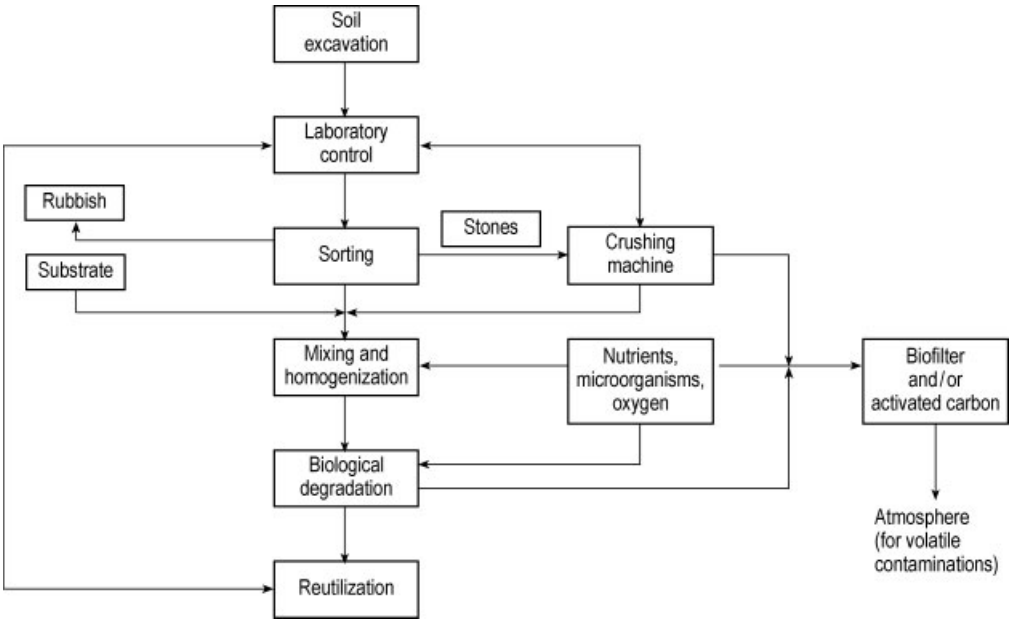

Figure 7. Schematic diagram of process for biological soil remediation

4.2. In-Situ Methods

In-situ methods are useful in the following cases:

Buildings, tanks, and roads on the ground must not be torn down. Consequently soil excavation is not possible.

The contamination has spread vertically and horizontally to such an extent that soil excavation is not sensible in economic and technical terms.

The most important prerequisite for in-situ remediation is a good permeability of the soil because the required substrates such as oxygen, adapted microorganisms, and nutrients must reach the contaminated area via water as the transport medium. As a rule, particularly in deeper soil layers the ambient conditions are substantially improved by setting up a rinsing circulation system. Furthermore, a practical arrangement of suction and injection wells ensures

that the treatment is limited to the contaminated grounds. Any undesired discharge of contaminants and bacteria to the surrounding area is thus avoided.

In addition to degradation in the soil itself, the installed water circuit also washes part of the contaminants out of the ground. Suitable measures, such as biological or chemical – physical water purification, further lower the contaminants concentration in the circulated water (Fig. 8).

4.3. Off-Site Methods

In the 1990s, soil remediation activities have extensively shifted from on-site methods at the respective contaminated sites to off-site treatment in stationary soil treatment plants (→ Soil, Chap. 7.1.2.1.).

Off-site the throughput of the soil remediation plants (total approximately 3.2×10^6 t/a), are dominated by biological methods. With approx. 2.3×10^6 t/a biological technology has surpassed soil washing by 1.4×10^6 t/a. Thermal plants reported a throughput of around 130 000 t/a [4].

In principle, the procedure and treatment in a biological soil remediation plant are comparable with on-site remediation.

5. Use of the Cleaned Soil

Soil that is cleaned by on-site remediation or at a stationary plant is again subjected to a meticulous chemical and biological examination. Besides the verification that the contaminations in the soil are below the legal threshold limits, a number of other quality characteristics are determined, including:

Grain mix
Humus content
Water storage capacity
Soil biology
Plant compatibility
Absence of weeds and
Hygienic safety

The results are important prerequisites for recycling of the cleaned soil. In addition, supporting toxicity tests are carried out with luminous bacteria, for example.

The entire process from reception of the cleaned soil to its transport to the site of reuse is constantly monitored and controlled by biological – chemical laboratories to ensure that no hazardous residual substances remain in the soil and the stipulated limit values are complied with.

Further biological treatment procedures in conjunction with on-site remediation are connected with cultivation measures at the respective sites. In a 30 – 40 cm thick layer the cleaned soil is placed on the fallow areas created where it is used as new topsoil. The new soil has a higher biological activity and an increased humus concentration than the usually depleted industrial fallow areas.

On the basis of planting measures suitable for the soil and balanced fertilization, the existing microorganisms can continue to decompose hydrocarbon products in the cleaned soil. In this way a residual concentration of 319 mg/kg could be minimized to less than 100 mg/kg.

This clearly indicates that contaminant degradation is not yet finished after successful conclusion of biological remediation. By restoration of the ecological functions of the soil, its fertility can be enhanced and the degradation of residual concentrations of contaminats can be reactivated. This degradation is possible up to a level that corresponds to the natural soil contents.

6. Case Examples

6.1. Remediation of a Tank Farm

For several decades gasoline, diesel fuel, and heating oil were handled at a tank farm more than 3000 m^3 in size in Saalfelden, Austria. Due to leakage, particularly as a result of corrosion, as well as during loading and unloading operations, mineral oil accumulated in the soil. In the course of investigations on the grounds, a vertical spread of contamination extending to approx. 12 m below ground level was determined. Analyses of soil samples showed a maximum concentration of hydrocarbons of 14 100 mg/kg. Since the contamination had penetrated down to the groundwater, the groundwater level was lowered for excavation of the soil masses.

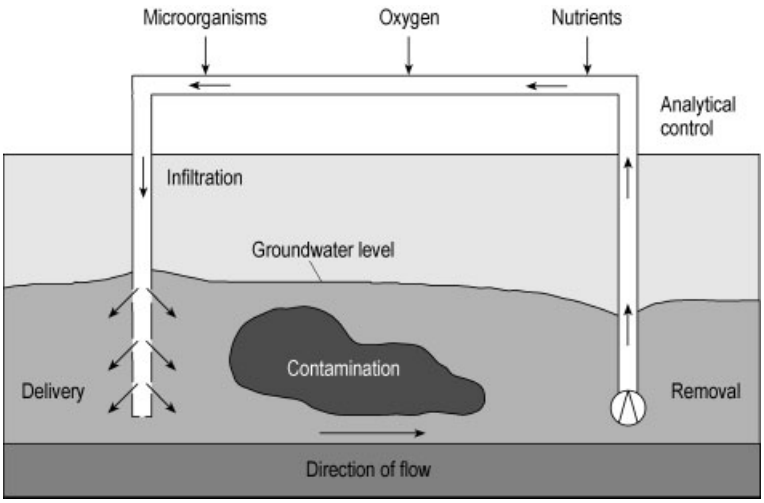

Figure 8. Microbiological and hydraulic in-situ remediation. According to [17].

Initial estimates indicated around 20 000 t of contaminated soil. In the course of excavation work, however, it turned out that the contaminated area was significantly larger. A total of about 110 000 t of soil were treated by means of microbiological methods. The total length of the large tents in which the soil was cleaned was over 1500 m. The conducted biotests indicated an expected degradation period of ca. 6 – 8 months, which was also confirmed in practice (Fig. 9). After remediation the entire soil material was used to refill the construction pit [18].

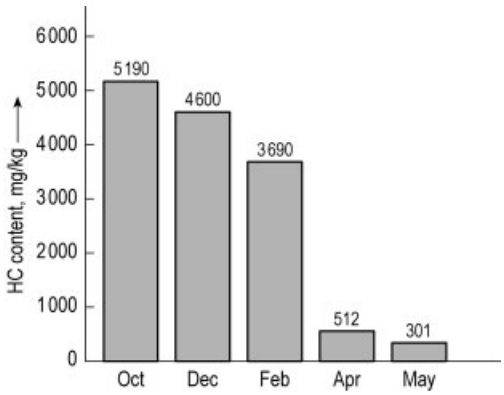

Figure 9. Degradation graph

6.2. Remediation of Soil Contaminated with TNT

A new humification method has been developed for the biological treatment of soil that is contaminated with highly toxic nitroaromatics such as TNT. This method can transform these compounds into nontoxic ones under sequentially reducing and oxidative ambient conditions [19], [20]. This is achieved by means of:

1) Targeted addition of specially combined, preconditioned substrates, and
2) Optimal mechanical soil tilling that is adapted to the respective degradation phases.

During treatment a gradual reduction of TNT takes place, resulting in the formation of triaminotoluene (TAT) as the final product (Fig. 10). Triaminotoluene is irreversibly immobilized in the humus matrix and thus detoxified. A first large-scale test, in which approx. 70 t of contaminated soil were treated, demonstrated the efficiency of the method.

Prior to the tests the soil was excavated on the grounds of a former explosives factory, screened to a grain size of 60 mm (the oversize portion was rejected), homogenized, and stored in closed transport containers. In a treatment tent the soil was intensively mixed with the substrate (consisting of several natural organic components). A total of 36 wt % substrate was added to the soil. The arrangement of the construction site is shown in Figure 11.

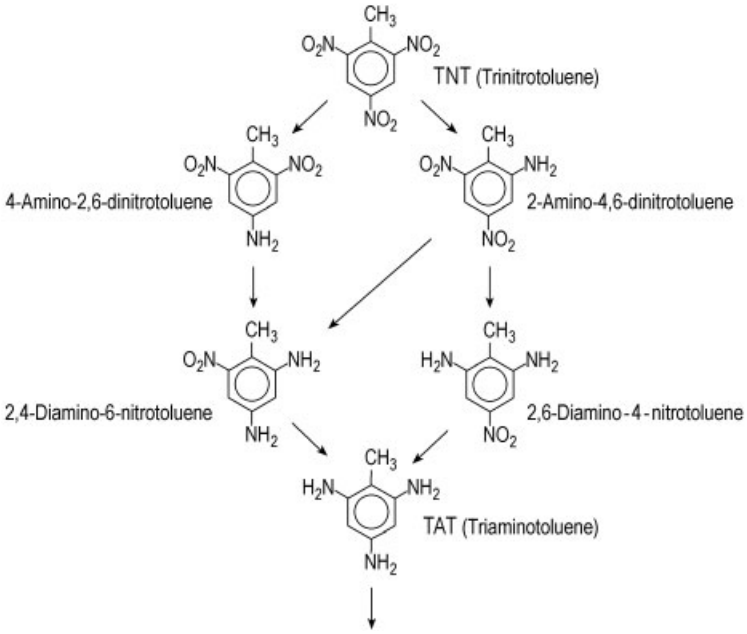

Figure 10. Transformation path of TNT

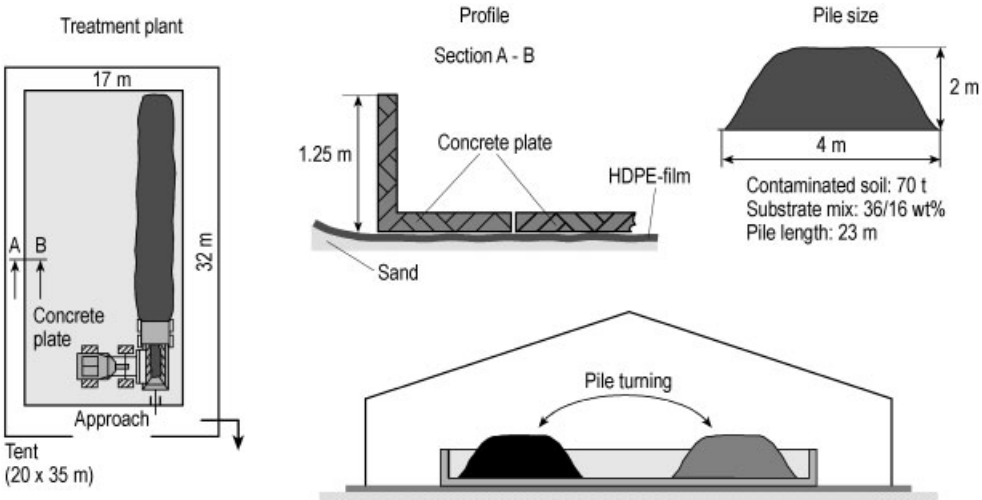

Figure 11. Arrangement of an on-site TNT treatment plant

The contaminant concentration was analyzed in an on-site laboratory. The original soil displayed a medium contamination with nitroaromatics of 1126 mg/kg. The initial contamination after addition of the substrate mix was around 699 mg/kg. The concentration of nitroaromatics during the first 14 days was within the detection limit range (0.5 mg/kg). On the 14th, 16th, and 18th test day nitroaromatics could still be detected in individual mix samples. In the subsequent investigations no contamination was detected up to the end of the test (86th day). Similar results are reported in [9], [20], [21]. On the basis of the positive results of the first test batch, the cleaned soil was placed back on the factory grounds.

7. References

General References

1. J. T. Cookson: *Bioremediation Engineering – Design and Application,* McGraw Hill, New York 1995.
2. R. E. Hinchee, F. Brockmann, C. M. Vogel (eds.): *Applied Bioremediation of Petroleum Hydrocarbons,* **vol. 3(6), from: Third International In-situ and On-site Bioreclamation Symposium, San Diego 1995,** Battelle Press, Columbus, OH 1995.
3. R. E. Hinchee, F. Brockmann, C. M. Vogel (eds.): *Microbial Processes for Bioremediation,* **vol. 3(8), from: Third International In-situ and On-site Bioreclamation Symposium, San Diego 1995,** Battelle Press, Columbus, OH 1995.

Specific References

4. H. J. Schmitz: "Die Jagd nach dem Boden war 1999 erfolgreich," *TerraTech* **4** (2000) 28 – 43.
5. W. D. Weißenfels et al.: "Adsorption of polycyclic aromatic hydrocarbons (PAHs) by soil particles: Influence on biodegradability and biotoxicity," *Appl. Microbiol. Biotechnol.* **36** (1992) 689 – 696.
6. R. E. Hinchee et al.: *Bioremediation of Chlorinated and Polycyclic Aromatic Hydrocarbon Compounds,* Lewis Publ., Boca Raton, FL 1994.
7. C. E. Cerniglia: "Microbial metabolism of polycyclic aromatic hydrocarbons," *Adv. Appl. Microbiol.* **30** (1984) 31 – 71.
8. O. Meyer et al.: "Development of techniques for the bioremediation of soil, air and groundwater polluted with chlorinated hydrocarbons: the demonstration project at the model site in Eppelheim," *Microb. Rel.* **2** (1993) 11 – 22.
9. H. Lenke et al.: "Biological treatment of TNT-contaminated soil. 2. Biologically induced immobilization of the contaminants and full-scale application," *Environ. Sci. Technol.* **32** (1998) 1964 – 1971.
10. T. Volk: Tom Volk's Fungus of the Month for August 1997, http://www.wisc.edu./botany/fungi/aug97.html.
11. D. Loske: Abbau von polycyclischen aromatischen Kohlenwasserstoffen durch Weißfäulepilze, Dissertation Göttingen 1991.
12. G. Schaefer et al.: "The effect of the addition of the white rot fungus Pleurotus ostreatus on the degradation of polycyclic aromatic hydrocarbons (PAH) in a long time contaminated soil," *J. Bioeng.* **9** (1993) 55.
13. K. L. Kreiton: Demonstration Bulletin, Fungal Treatment Bulletin, USEPA-RREL/USDA-FPL, EPA/540/MR-93/514, United States Environmental Protection Agency, 1993.
14. W. Fabig: "Mikrobiologie und Chemie bei Verunreinigungen von Boden und Grundwasser mit Kohlenwasserstoffen", in W. J. Bartz, E. Wippler (eds.): *Angewandte Mikrobiologie der Kohlenwasserstoffe in Industrie und Umwelt,* **vol. 164, (Reihe: Kontakt und Studium),** expert-Verlag, Ehningen 1998, p. 37.
15. R. M. Atlas: "Microbial Degradation of Petroleum Hydrocarbons: An Environmental Perspective," *Microbiological Reviews* **45/1** (1981) 180 – 209.
16. O. Meyer et al.: "Pilot Plant Stage Bioremediation of CKW- and BTEX Contaminated Soil by In-situ Infiltration in Combination with On-site Water and Air Treatment at the Model Site Eppelheim", in W. J. van den Brink et al. (ed.): *Contaminated Soil,* Kluwer Academic Publ., Dordrecht 1995, p. 843.
17. Landesamt für Umweltschutz Baden-Württemberg (ed.): Materialien zur Altlastenbearbeitung, *Handbuch Mikrobiologische Bodenreinigung,* **vol. 7,** Stuttgart 1995.
18. G. Vollmer-Heuer: "Bodensanierung eines Mineralöltanklagers in Saalfelden," *Umweltschutz* **3** (1991) 13 – 14.
19. U. Walter et al.: "Anaerob/Aerob-Verfahren zur biologischen Behandlung sprengstoffkontaminierter Böden," *TerraTech* **3** (1998) 42 – 45.
20. J. Michels: "Processes for the bioremediation of soil," Compilation of current projects of the Joint Research Group, Federal Environmental Agency, BMBF, Bonn 1999.
21. D. Bruns-Nagel et al.: "Bioremediation of 2,4,6-trinitrotoluene-contaminated soil by anaerobic/aerobic and aerobic methods," **vol. 4(2) from: Fourth International in Situ and On-Site Bioremediation Symposium,** Battelle Press, Columbus, OH 1997, pp. 9 – 14.

Chemical Conservation of Stone Structures

ROLF SNETHLAGE, Bayerisches Landesamt für Denkmalpflege, München, Federal Republic of Germany
EBERHARD WENDLER, Fachlabor für Konservierungsfragen, München, Federal Republic of Germany

1. Introduction 1
2. Traditional Stone-Conservation Products 3
2.1. Limewater and Lime Wash 3
2.2. Linseed Oil 3
2.3. Waxes 4
2.4. Fluorosilicates 4
2.5. Water Glass 4
3. Organosilicon Compounds 5
3.1. Strengtheners 5
3.1.1. Tetraethyl Orthosilicate 6
3.1.2. Flexible Gels 8
3.1.3. Coupling Agents 8
3.1.4. Effect on Stone Properties 8
3.2. Hydrophobic Products 9
3.2.1. Formulations 9
3.2.2. Combined Products 12
3.2.3. Microemulsions 12
3.2.4. Effect of Hydrophobation on Stone Properties 13
3.2.5. Durability 14
4. Acrylates 15
4.1. Reactive Acrylate Systems 15
4.2. Polymer Acrylate Systems 16
4.3. Total Impregnation with PMMA . . 17
5. Epoxy Resins 18
6. Polyurethanes 20
7. Polyester and Alkyd Resins 21
8. Polymer Dispersions 22
9. Conclusions 25
10. References 25

1. Introduction

Since early antiquity, the deterioration of building stones has been a well-known problem in the preservation of our cultural heritage. The Great Sphinx in Giza, which had been covered by sand for centuries, was excavated and restored for the first time by THUTMOSIS IV (1400 B.C.). The vestiges of this intervention, revealing several layers of square stones which had been put in place to repair the deteriorated surface of the body of the Sphinx, can still be seen today. The cathedral stonemasons workshop – some of them existing since the erection of the cathedral – demonstrate the necessity of steady maintenance and repair. According to the opinion of most scientists, the rate of damage to historical monuments has increased exponentially since the beginning of industrialization in the 1800s [1], [2]. In fact, many photographs taken around 1900 seem to confirm the enormous progress of deterioration in monuments in the past 100 years (Fig. 1).

However, the state of many monuments was already alarming in the 1800s. Gypsum copies which were produced from the sculptures of the Fürstenportal at Bamberg cathedral between 1904 and 1906 clearly demonstrate that at the turn of the century the degree of deterioration of the 24 figures was almost the same as in 1990 (Fig. 2).

Because there is only insufficient documentation on the state of monuments, our knowledge is still too limited to decide whether increased deterioration has taken place during the course of the 1900s. In many cases the destruction cannot be attributed to air pollution effects alone, but is also the result of defective or inappropriate maintenance.

Historical monuments reflect the development of air pollution during their life history. The damages that we observe today are the consequences of a long-term accumulation of damaging substances and growing damage factors. Exposure programs started during the past 25 years show a correlation between material loss and intensity of air pollution [3–8]. Several authors estimated damage functions for SO_2 and H^+, especially for calcareous stones [3], [9–12]. A synergistic effect of SO_2 and NO_x on the damage rate was found [13], [14]. The susceptibility of different stones to various gaseous air pollutants becomes evident in the different deposition rates for SO_2, NH_4^+, and NO_x, which depend on the physical and chemical properties and the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.d06_d01

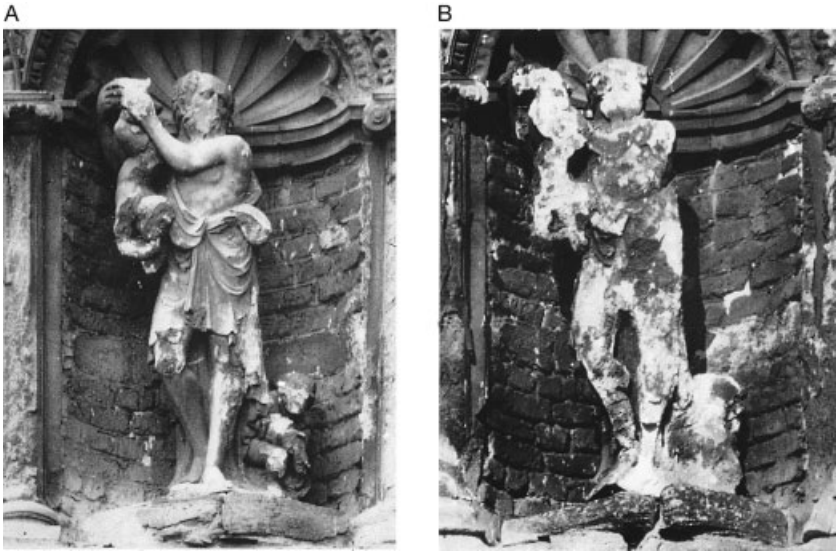

Figure 1. Photographs of a sculpture on the facade of Castle Horst in North Rhine-Westphalia, Germany, taken in 1913 (A) and 1983 (B). The comparison demonstrates the progress of deterioration during 70 years of exposure.

moisture content of the stone surface [15–19]. Both parameters are very much influenced by the formation of biofilms, which enhance the deposition of particulate matter and gaseous pollutants [20–22]. Further information regarding the correlation of air pollution and stone deterioration can be found in [23], [24].

Scientific investigation into the deterioration and conservation of building stones began in the 1800s and depended on the progress in chemical research. In particular, conservation was mainly concerned with the preservation of the many works of art that entered European museums after having been excavated in antique sites around the Mediterranean Sea [25]. Comprehensive bibliographies on the use, deterioration, and conservation of natural stone were presented in [26], [27]. An interesting account on former treatments in England is given in [28], which concluded that most conservation products had failed or showed neither positive or negative effects. Like today, the products that were used for stone conservation were primarily developed for other purposes and then recommended for conservation in order to open up additional sales. The list of historic stone-conservation products is quite limited and comprises materials which were well-known since antiquity or invented in the 1800s. Probably the most complete compilation of commercially available stone conservation products can be found in [29], in which producers, trade names, and chemical compositions of the products are described. In the following sections the most important methods and materials for stone conservation are described.

Figure 2. Three apostles of the Fürstenportal at Bamberg cathedral in 1990.

2. Traditional Stone-Conservation Products

2.1. Limewater and Lime Wash

Limewater is a saturated solution of $Ca(OH)_2$. It contains 0.18 g of dissolved $Ca(OH)_2$ per liter. Brittle stones are soaked with limewater several times, a procedure that takes weeks or even months. The idea is to replace the dissolved binding cement of stones by introducing new calcium carbonate according to the reaction

$$Ca(OH)_2 + CO_2 \longrightarrow CaCO_3 + H_2O$$

However, a strengthening effect has never been proved, because the total amount of deposited $CaCO_3$ is very small and the crystals are mainly formed near the surface where the $Ca(OH)_2$ can react with atmospheric CO_2. In fact, treatment with limewater has rather negative effects due to the enormous quantities of water that are applied to the stone. In addition, the alkalinity of the limewater increases the rate of SO_2 deposition, so that all of the $CaCO_3$ formed is converted into gypsum within a very short time.

Lime washes are thin layers of very fine lime mortar. Their thickness is on the order of a few millimeters and should not exceed 5 mm. The purpose is to protect the original stone surface from incident rain and air pollutants by means of a sacrificial layer. When the sacrificial layer is worn down it can be renewed by a new layer. The method has attracted increasing interest in several European countries, especially regarding the criteria of compatibility and re-treatability [30]. It is also quite popular in the United Kingdom, where most of the historic monuments and cathedrals are built of oolitic limestones, so that the use of lime-wash mortars has a greater tradition than elsewhere [31]. One of the most important examples of applying lime washes is the treatment of the figures in the west facade of Wells cathedral that was carried out during the restoration of the whole facade in the 1980s. The lime wash, which has the consistency of a putty, is rubbed into surface cracks and delaminating flakes with soft brushes, thus stabilizing loose surface parts.

Tests on sample areas on pinnacles of Cologne cathedral showed an insufficient durability of the lime washes, which were completely weathered after only one year of exposure [32]. Results obtained in Austria demonstrate that in order to achieve the desired effectiveness lime washes must be modified with polymer dispersions [30] or even hydrophobized, so that their properties can no longer be strictly distinguished from those of dispersion paints.

2.2. Linseed Oil

Linseed oil, a mixture of oleic, linoleic, and linolenic acids, (→ Fats and Fatty Oils, Chap. 13.2.4.3.) belongs to the so called drying oils (→ Drying Oils and Related Products). Drying occurs through reaction of the C=C double bonds to form a polymer.

Linseed oil is commercially available in different grades according to the degree of refining. Stand Oil is produced by heating linseed oil in the absence of air in closed vessels (→ Drying Oils and Related Products, Chap. 8.2.). By this process, several of the reactive double bonds react with each other, thus forming larger molecular units. The reactivity and ability to form a polymer network decreases. The drying is enhanced by lead compounds. For this reason linseed oil is the typical binder in oil paints and in particular in lead white paints.

Linseed oil has been in use for stone preservation for centuries, probably even since antiquity. It was the main constituent of many products that were on the market from the 1800s up to ca. 1950 [29]. It has been applied to limestones and porous sandstones. Ref. [33] underlines both its strengthening and hydrophobic effect, moreover arguing that even its swelling due to the action of moisture could be advantageous because further penetration of rainwater would be eliminated. However, its insufficient penetration depth, discoloration, and steadily increasing brittleness together with decreasing hydrophobic effectiveness are detrimental to its use.

2.3. Waxes

For centuries, natural waxes have been used for protecting and polishing susceptible stone surfaces. The most frequently used wax is bees wax which is available in different grades with increasing degrees of refinement. It is still applied

today, e.g., for polishing marble tombstones on the inner and outer walls of churches. Bees wax possesses the advantage of not becoming brittle even after decades, but its stickiness causes ugly dirt depositions, especially in the open air. Modern stone-care waxes are based on various paraffins and are sold in dissolved (organic solvents) or dispersed (in water) state.

2.4. Fluorosilicates

Hexafluorosilicate salts of the cations Mg, Al, Zn, Pb, Cu, Cr, Fe, ammonium ion, and alkali metals are still applied, mainly to improve the density of pores and the chemical resistance of concrete. The effect is apparently based on the deposition of amorphous silica gel. In the literature, two different reaction varieties can be found which describe the chemical reactions in case of concrete or sandstone.

Concrete:

$$2\,Ca(OH)_2 + MgSiF_6 \longrightarrow 2\,CaF_2 + MgF_2 + SiO_2 + 2\,H_2O$$

Sandstone and limestone:

$$2\,CaCO_3 + MgSiF_6 \longrightarrow 2\,CaF_2 + MgF_2 + SiO_2 + 2\,CO_2$$

Although the formation of amorphous silica is not doubted, both reactions are tentative because the reaction products CaF_2 and MgF_2 were never clearly determined. An effect can only be expected for limestones or calcareous sandstone.

Mg- and Zn-fluorosilicates were very wide spread after their synthesis in 1883 by Kessler. Whereas ref. [25] favors their application for limestones, ref. [33] objects that the strengthening would be restricted to the very surface and could yield negative results in the long run. Because the salts are strongly acidic, the same discolorations can occur as with the use of hydro fluoric acid.

2.5. Water Glass

Several patents have been issued for the synthesis and production of stone-conservation products based on water glass [26]. Early water glass paints were developed by Keim [34].

Water glass has a SiO_2/M^+-ratio of 4 : 1 (M = Na, K). In aqueous solution the silicate anion is polymerized into three-dimensional, nearly spherical particles which consist of ca. 90 SiO_2 units linked by oxygen bridges [35]. In its interior, the particles are free of OH groups and have a structure similar to that of β-cristoballite. On the surface, each Si atom is bound to an OH group. At pH 12, ca. 40 % of the OH groups are ionized to O^- groups. The potassium or sodium ions are hydrated by water molecules.

The reaction of water glass to form amorphous silica is mostly explained by the reaction with atmospheric CO_2 according to the equation:

$$K_2SiO_3 + CO_2 + n\,H_2O \rightarrow SiO_2 \cdot n\,H_2O + K_2CO_3$$

The reaction, however, is very slow due to the small amount of CO_2 in the air. Therefore, another process is mainly effective. The silica gel is very quickly formed in contact with porous materials when the water soaks into the capillaries and is withdrawn from the solution. Consequently, the solution is increasingly concentrated until the negatively charged colloidal silica particles come into such close contact to each other that the OH groups at their surfaces condense and macroparticles are formed, which then precipitate from solution [35].

This also explains why water glass is not appropriate for consolidating natural stones. Because the silica component is retained at the surface, only water penetrates deeper into the stone. The strongly consolidated surface zone rarely exceeds 1 or 2 mm. It is at risk of spalling off because it is underlain by an unconsolidated soft stone layer. Another reason for excluding water glass from conservation is the high content of alkali metals, which contaminate the treated material with hygroscopic alkali metal carbonates that can cause severe salt damage.

Silica sols gained importance in conservation only recently, but their application is described here because their chemical constitution is very similar to those of water glasses. Silica sols are solutions of colloidal silica (→ Silica, Chap. 4.). The particle size (7 – 125 nm) is greater than that of the silica particles in water glass solutions (2 – 3 nm). The particle size depends on the amount of dissolved silica, which ranges from 15 to 50 wt %. Silica sol solutions are slightly cloudy or milky depending on the size of the particles. The main advantage of using silica sols is

the greatly diminished content of alkali metal ions, mostly sodium, which reaches 0.5 wt % at most. Silica sols are produced from Na silicate solutions by reducing the sodium content by ion exchange with H^+ ions.

As in the case of water glass, the precipitation of the silica gel is caused by the loss of water, so that the hydroxyl groups condense and larger particles are formed. Although the particle size of some formulations of silica sols is very small, the product is not appropriate for consolidating porous materials because the silica particles are retained on the surface, where they tend to form an impermeable layer. The main field of application is the production of stone-repair mortars for filling smaller cracks or reshaping missing edges of square stones (Fig. 3). Thin silica-sol-bound washes are also used to protect surfaces with a sacrificial layer. Because of the low alkali metal content, the treated stones are not contaminated with salt and the risk of efflorescence is negligible. Detailed recipes for producing silica-sol-bound repair materials are described in [36].

Figure 3. Detail of the greensandstone masonry of St. Peter's church in Straubing, Bavaria, with application of silica-sol-bound repair mortars

3. Organosilicon Compounds

3.1. Strengtheners

During the weathering process, the loss of binding material leads to a decrease of strength in the outermost zone of sandstones and other building stones. The idea of structural strengthening is to replace the lost binding material by introducing an artificial binder into the space between the disintegrated and loose mineral grains. This can be achieved by applying reactive liquids which are taken up by the capillarities of the material and subsequently deposited a solid polymer or gel that fixes the loosened grains together. The aim should be a homogeneous depth profile of strength, as in the fresh unweathered stone.

Several polymers and gel-forming liquids are able to enhance the strength of a deteriorated stone. Due to its chemical similarity, a silicate-like system should be favored. The disadvantages of water glass are discussed in Section 2.5. Therefore the use of silicon organic compounds seems promising, since the gels formed from them are free of other compounds which might reduce the durability of the measure.

3.1.1. Tetraethyl Orthosilicate

Ethyl silicate has been widely used as stone strengthener since the late 1960s [37]. It is prepared from silicon tetrachloride by solvolysis with ethanol/water mixtures [38]. Depending on the amount of water added, a part of the resulting tetraethyl orthosilicate (TEOS) forms larger units by condensation. Monomeric or oligomeric TEOS is hydrolyzed by water present in the pore space of stones. Two hydrolyzed silanol (Si–OH) groups can condense by intermolecular release of a water molecule, thus forming stable siloxane (Si–O–Si) bonds. The final result is a solid gel structure which interacts with the mineral components of the stone (e.g., quartz, feldspars, rock fragments), whereby in many cases a stable contact results (Fig. 4).

Methyl silicate (tetramethyl orthosilicate TMOS), which reacts much faster than TEOS and was therefore frequently used in the 1960s and 1970s, is no longer used because it releases toxic methanol on hydrolysis.

The liquid TEOS products are applied either free of solvent or in ketone or alcohol solvents. Under the action of water in the pore space of

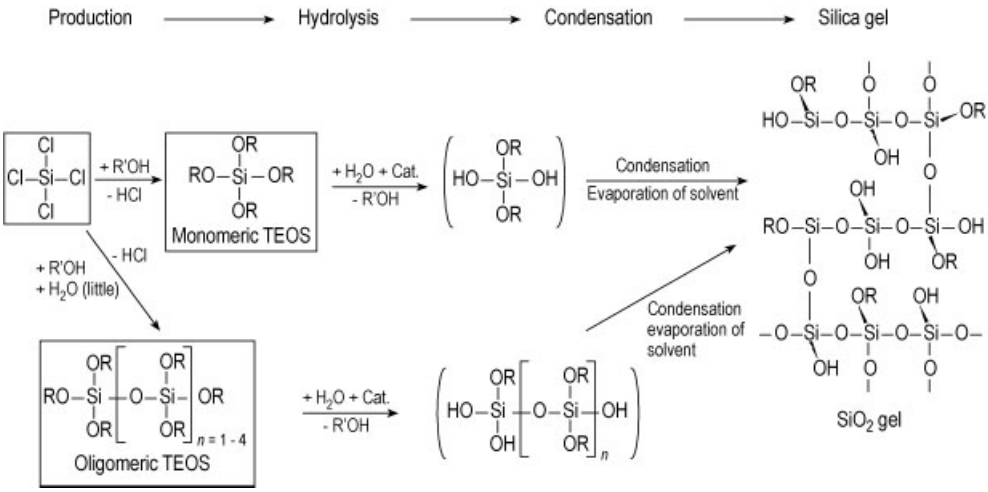

Figure 4. Formation of TEOS from $SiCl_4$. Gel formation by hydrolysis and subsequent condensation.

stones, a macromolecular gel is formed by hydrolysis, alkaline cleavage of a proton (Eq. 1) and subsequent condensation (Eq. 2). The overall reaction is catalyzed either by acids or bases, in the latter case yielding more branched rigid gels, while linear segments form in the presence of acids [39].

$$(RO)_3SiOH + OH^- \longrightarrow (RO)_3SiO^- + H_2O \quad (1)$$

$$(RO)_3SiO^- + (RO)_3SiOH \longrightarrow (RO)_3Si\text{–}O\text{–}Si(OR)_3 \quad (2)$$

In presence of bases, condensation is the rate-determining step. Step (1) becomes faster with increasing acidity of the proton. Since larger oligomers containing many Si–O–Si siloxane bridges are weaker bases than smaller units, their protons react more slowly. For this reason, the oligomers preferentially react according to Eq. (1), finally yielding highly branched molecules.

In the presence of acid catalysts the condensation of hydrolyzed species is faster than the hydrolysis itself, which then becomes the rate-determining step. In this case, hydrolysis is initiated by the electrophilic attack of a proton on an oxygen atom of the alkoxyl group (Eq. 3):

$$(RO)_3Si\text{–}O\text{–}R + H^+ \longrightarrow (RO)_3Si^+ + R\text{–}OH \quad (3)$$

This reaction does not depend on electronic but on steric effects: end groups are hydrolyzed faster than alkoxyl groups in the middle of a siloxane chain. Therefore, long linear chains are formed under acidic conditions.

Commercial products for stone consolidation are usually catalyzed by organotin compounds (e.g., dibutyltin dilaurate) or amines, which give a slower gel-forming rate and thus allow deeper penetration. The TEOS product most frequently used is Stone Strengthener OH or Conservare OH, a mixture of three parts of monomeric and oligomeric TEOS units dissolved in one part of solvent (mainly methylethyl ketone).

Depending on the reaction rate, which is influenced by temperature, humidity, and catalyst, different gel structures may form [40], [41]. In the course of hydrolysis and condensation which are competing steps, a network of connected ring structures $(\text{–O–Si–O–Si})_n$ is formed. Macroscopically, the gel consists of spherical particles several nanometers in diameter, which are piled together in an open structure with cavities between them. The rigidity of the three-dimensional structure prevents the condensation of all remaining silanol groups because the distance to neighboring silanol groups is too far for forming a siloxane bond. Moreover, at room temperature, the cavities the network are filled with water, ethanol, and TEOS molecules [42]. For steric reasons, the reaction will never lead to a stoichiometric SiO_2 network. The maturing process in the gel structure continues for several years and is accompanied by a continuous loss of water, which may lead to shrinking gaps in the structure as the aging of the gel proceeds (Figure 5).

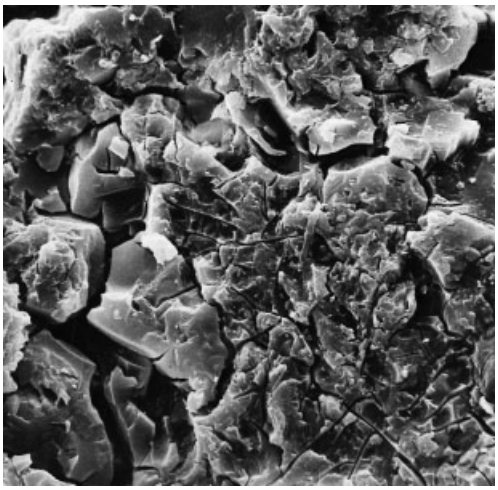

Figure 5. Gel in the pore space of Lechbruck sandstone treated with Stone Strengthener OH. SEM image: 150 µm × 150 µm

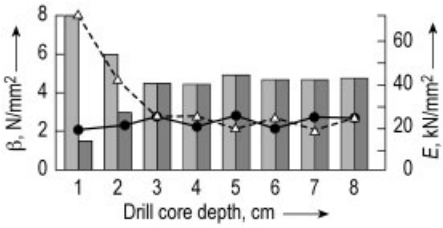

Figure 6. Bars: Depth profiles of the flexural strength β of a weathered sandstone core (light grey) and a core treated with Stone Strengthener OH (dark grey). Lines: corresponding depth profiles of the elastic modulus E, determined from the slopes of the stress – strain curves. Greensandstone, Alte Pinakothek, München.

Thermogravimetric/differential thermal analysis (TG/DTA) measurements to monitor the mass loss during gel aging have been carried out [40], [43]. Detailed microscopic investigations on the interface between mineral surfaces and SiO_2 gel in the pore space of stones are given in [44].

The mechanism of fixing of the gel to the mineral substrate are not yet fully understood. The surface of quartz grains bears silanol groups which may well act as anchor groups for the gel. Thus, covalent siloxane (Si–O–Si) bonds can be formed by condensation at the interface. Hydrogen bonding between silanol groups of the gel and those of the substrate may also lead to a sufficiently strong but reversible adhesion.

3.1.2. Flexible Gels

Since a few years, modified TEOS products forming flexible gels have been commercially available. Gels formed from these products do not form shrinking gaps during the aging process. Flexibility of TEOS can be achieved by integrating linear soft segments into the rigid network of silica gels [45]. In this way, thin elastic and flexible films are obtained which cover the surface of macropores, binding together the mineral components. Compared with normal gels, the durability of the modified gels should be improved. The resulting gels can be water-repellent or hydrophilic, depending on the type of soft segment.

Linear segments of poly dimethylsiloxane can be integrated into the growing gel structure, and function like "expanders" that impart the desired elasticity to the gel. Due to the methyl groups in the bridging segment, these gels show water-repellent properties. Hydrophilic gels with elastic properties can be obtained by using polyether groups as bridging segments.

3.1.3. Coupling Agents

In most cases, sandstones show a marked increase in strength when treated with TEOS products. However, there are still major problems concerning the consolidation of limestones. The reason for failure seems to be an insufficient binding of the gel to the carbonate surface [46]. Moreover, the condensation reaction is retarded in the presence of calcite but not in the presence of quartz [47].

The effectiveness of TEOS in the consolidation of limestones can be improved by using adhesive coupling agents as additives [48]. Phosphate compounds can covalently link to the growing gel structure during the condensation reaction. The negatively charged phosphate groups are oriented towards the mineral surface, since calcite has a positive zeta potential in the normal pH region. Due to electrostatic interaction, the gel is tightly fixed to the mineral components. In contrast to the use of Stone Strengthener OH, a phosphate-modified TEOS product leads to a sufficient increase in strength when applied to Savonnière limestone.

In general, trialkoxysilanes with functional groups of different polarity (depending on the type of substrate) are useful coupling agents since they are covalently integrated into the gel network structure. The integration of aminopropyl trialkoxy silane (APS) into the gel network results in improved bonding to quartz and mica surfaces [48]. The interaction of the silanol groups of APS with glass surfaces is described in [49].

3.1.4. Effect on Stone Properties

Figure 6 shows an example for the depth profile of strength, determined on drill-core slices of a greensandstone core treated with Stone Strengthener OH compared to a drill core of untreated, weathered stone. The modulus of elasticity is enhanced simultaneously with the increase of strength, while it decreases with the structural disintegration caused by weathering.

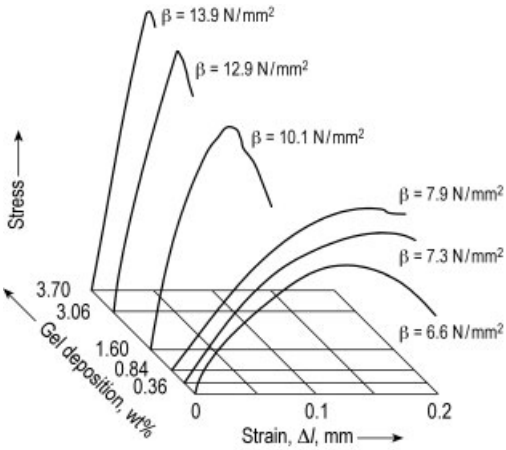

Figure 7. Stress – strain curves of the biaxial flexural strength of Schilfsandstone slices. The slices were treated with different concentrations of a TEOS product. Front: untreated. The slope of the curves in the linear region (Hook region) is proportional to the E Modulus.

Investigations clearly demonstrate a non linear correlation between the amount of gel deposition and the modulus of elasticity E (Fig. 7) [50]. Low gel deposition values only lead to a negligible increase in E. However, if the gel deposition exceeds a certain value, a sharp increase in E is observed. Thus, in case of a high gel deposition, the stress – strain behavior of the outermost, treated zone may be completely different from that of the untreated interior. Especially when the penetration depth is low, shearing forces between the treated and the untreated zones are generated when the external humidity or temperature changes. It is often argued that the shearing forces may lead to subsequent delamination of surface scales.

As expected, the increase in the modulus of elasticity is diminished when flexible gels are deposited in the pore space of sandstones. Figure 8 shows a correlation between the E-modulus and biaxial flexural strength β, determined on Schilfsandstone slices treated with commercial Stone Strengthener OH and modified TEOS product in comparison with the untreated material. The increase in flexural strength is comparable in both cases, but the E modulus is lower for treatment with the modified product.

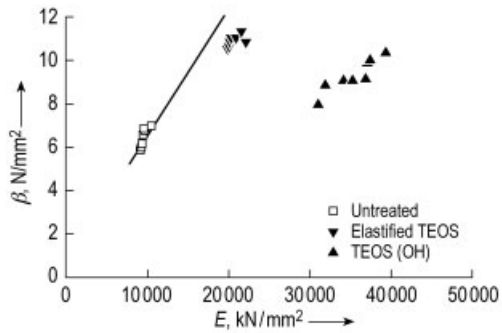

Figure 8. Correlation between E modulus and biaxial flexural strength β for untreated and consolidated stone (yellow schilfsandstone from Maulbronn).

As shown above, silica gels formed from ethyl silicate tend to form shrinking fissures due to continuous loss of structural water. However, laboratory and field studies [40] on numerous treated sandstones clearly show that there is no loss in efficiency even after 25 years. However, in the presence of hygroscopic salts, the durability may be greatly diminished due to mechanical stresses exerted by dissolution and recrystallization processes [51], [52].

TEOS products can be also used as binders for other restoration materials such as repair mortars, refilling grouts, washes, and varnishes [36], [53]. Compared with lime mortars, which are often used as sacrificial layers, TEOS products have the advantage of better durability against rain and acidic air pollutants. A further advantage is that full compatibility is guaranteed

when all restoration materials are based on the same chemical system.

3.2. Hydrophobic Products

The purpose of a hydrophobic treatment is to reduce the amount of capillary water taken up by a porous material. Since water participates in most damage processes (e.g., dissolution of calcite, ingress of reactive acidic air pollutants, swelling, frost), hydrophobization seems a useful concept for diminishing the weathering rate of stones. In ideal case, a water drop does not spread on a treated surface but remains in a sphere-like shape with a contact angle of $> 120°$ for several hours until it evaporates (Fig. 9).

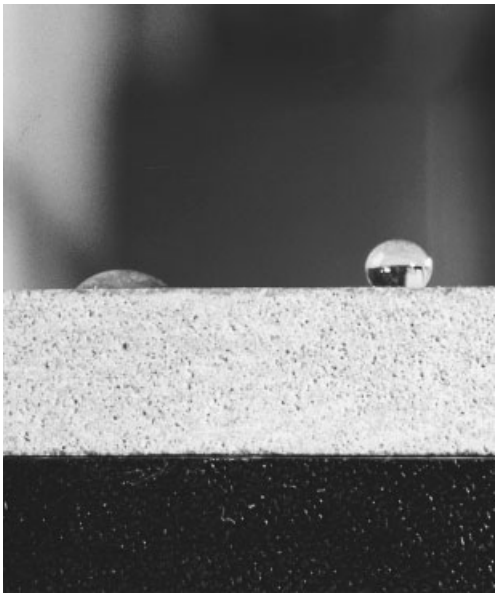

Figure 9. Sperical water drop on a treated sandstone surface (right) in comparison to an untreated sandstone, where the drop spreads and moistens the surface (left)

3.2.1. Formulations

Nearly all water-repellent agents used for the hydrophobic treatment of stone form silicone resin films as a final product. The precursors may be monomers like alkyltrialkoxysilanes or oligomeric/polymeric siloxanes. These compounds are frequently applied in aromatic hydrocarbon (rather than ethanolic) solutions of various concentrations. Similar to TEOS, after hydrolysis and condensation the precursors are linked together by siloxane bonds, finally forming a polymeric silicone resin (polysiloxane). Solutions of silicone resins can also be used. After evaporation of the solvent, the resin remains in the structure as a water-repellent film.

The basic compounds for the production of alkyltrialkoxysilanes as well as their oligomeric and polymeric homologues are trichloroalkylsilanes. The monomeric alkyltrialkoxysilanes are formed by addition of alcohol, while alcohol/water mixtures result in the formation of oligomeric siloxanes. A larger amount of water leads to a longer chain. Polymeric silicon resins are obtained by reaction with pure water (Fig. 10).

In contrast to alkylsilanes/siloxanes, the siliconates formed by addition of alkali metal hydroxides (Fig. 10) are water-soluble due to their ionic properties. This may be advantageous in some applications. However, since soluble carbonates are formed as undesired byproducts, the use of these products for hydrophobic treatment of building facades has decreased. Siliconates, however, are still often applied as injection liquids against rising damp.

Unlike the gel formation of TEOS, where each silicon atom has four oxygen neighbors, in alkylsilanes/siloxanes one binding position is blocked by the alkyl group. Thus, sheetlike structures are formed during condensation.

As in the case of TEOS gels, the binding mechanisms of the polysiloxane films to the mineral surface are not fully understood. Since hydrolysis and condensation are competing steps, a growing siloxane still has silanol groups which may interact with OH groups on the mineral surface. The binding interaction may be achieved by hydrogen bonding or by subsequent covalent bonding after release of water (Fig. 11).

In the ideal case, the alkyl groups ("umbrellas") should be oriented outwards thus giving an optimized water-repellency to the coated surface (Fig. 11). However, due to the different interatomic distances of silanol and other OH groups on the surface of minerals, which are different from those of the silanol groups in the forming siloxane, it is most improbable that sheath structures can develop on the surface. In addition, the formation of sheets is prevented by steric

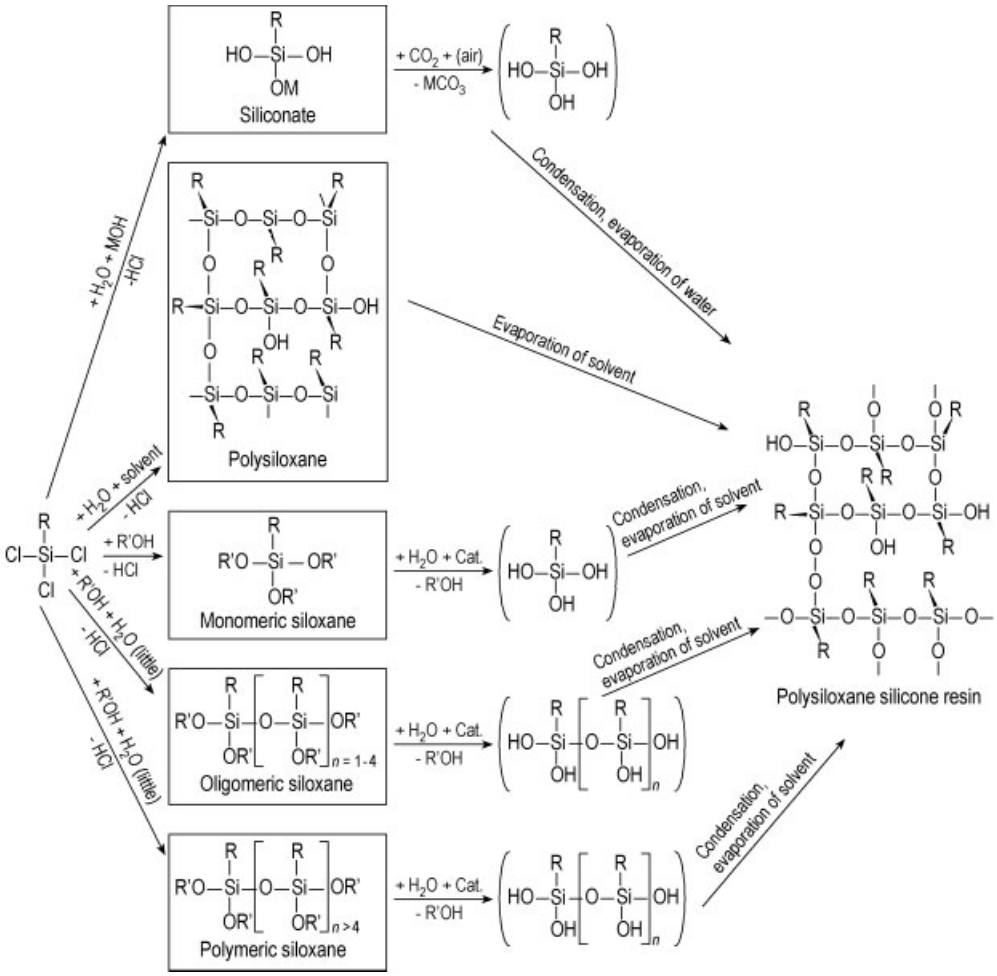

Figure 10. Production of different silicon organic hydrophobic agents from trichloroalkylsilane. Reaction steps required to form a silicone resin film in the pore space of stones [54].

effects of the bulky alkyl groups. Therefore, silicon resin films more likely consist of tangled R–SiO_3 chains [46].

Since the concentration of silanol groups on the surface of quartz is rather high, sandstones with a high quartz content show good hydrophobic properties when treated with silanes/siloxanes. The films are tightly connected to the mineral surface. Similar results are obtained when the surface of a stone has been previously coated with gel by a TEOS strengthening treatment. This seems reasonable, since silica gel has a high concentration of free silanol groups. In contrast, clay minerals offer silanol groups to which siloxanes can link only along their edges, so that polysiloxane films are suspended like sails from edge to edge of different clay particles (Fig. 12). Since the density of silanol groups on the basal planes of clay minerals is negligible, the interaction is much weaker in these areas. Probably, van der Waals interactions are exclusively responsible for the contact between clay basal planes and polysiloxane films. Therefore, water molecules can displace the silicone resin film in the persistent presence of water. In the case of calcite, it is likely that the weak Ca–O–Si bonds which link the calcite surface to the polysiloxane film can be easily hydrolyzed and broken up by protons.

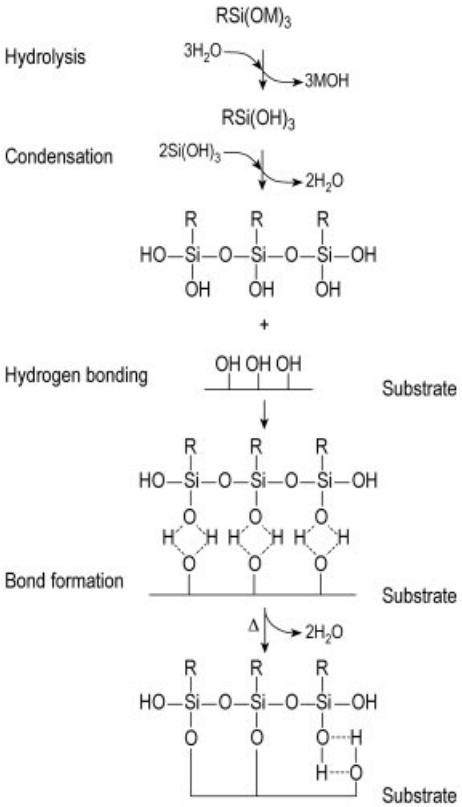

Figure 11. Reaction and bonding mechanisms of alkyltrialkoxysilanes [55].

Figure 12. Holey, house-of-cards-like chlorite structure in a clay-rich sandstone, spanned over by a silicone resin film by a treatment with a polysiloxane (W 090 S). Breitbrunn sandstone. SEM image: 44 × 44 µm [46].

The water-repellent properties of a silicone resin are determined by the length and degree of branching of the alkyl groups. While products with long alkyl chains show better water-repellency, their ability to penetrate into narrow pores is reduced. While this effect is of major importance in the case of concrete, most sandstones show a sufficient penetration of products with long alkyl chains since their capillary pores are in the range of 10 to 100 µm, which exceeds the diameter of a bulky alkyl group by a factor of around 100 to 1000.

Under basic conditions (concrete, fresh mortar), the Si–C bond can be cleaved easily when the alkyl chain is short, and methylsilicone resin is not stable under these conditions. Hydrophobing agents showing stability towards alkaline attack consist of isobutyl or isooctylsilanes/siloxanes.

3.2.2. Combined Products

Since the early 1970s, combined products have been available consisting of mixtures of TEOS (strengthener) and oligomeric siloxanes (hydrophobing component). The main advantage is that both strengthening and hydrophobing can be carried out in one step, which makes the planning of restoration measures easier and reduces the cost. Moreover, the interaction of the gel surface with the silicone film is optimized when both macromolecules grow at the same time (cross-condensation). The most prominent of such conservation products is Wacker H, a mixture of TEOS and oligomeric siloxanes with different alkyl groups such as methyl, butyl, and isooctyl.

3.2.3. Microemulsions

Since 1990, aqueous hydrophobic agents are available as microemulsions which have the main advantage of being free of toxic solvents [56]. They contain droplets of a silane/siloxane dispersed phase stabilized by two types of surfactants. Since the surfactants also consist of silanes/siloxanes, the system only contains the reactants necessary to form the desired polysiloxane film inside the pore space.

In normal oil/water-type emulsions, the disperse phase is covered by a single surfactant

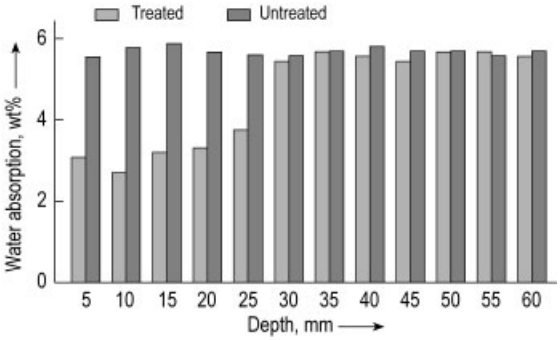

Figure 13. Depth profile of water absorption, determined on slices from a treated and an untreated sandstone core by weighing after 24 h of immersion in water. Bunter sandstone from Fritzlar Cathedral, treated with W 290 S (oligomeric siloxane) after strengthening with Stone Strengthener OH.

that forms a molecular layer with quasicrystalline properties ("solid condensed film"). Due to strong cohesion forces within the disperse phase, these well-ordered structures are rigid and exhibit only little tendency to form curved surfaces. Thus, the colloidal spherical particles have a rather large diameter of about 1000 nm, and these emulsions are not capable of entering the narrow capillaries of the stone structure.

Using an additional second type of surfactant (cosurfactant) with a shorter nonpolar alkyl group disturbs the ordering of the monomolecular surfactant layer. As a consequence, the cohesion forces of the disperse phase exceed the attractive forces between the two surfactant phases, so that the tendency of the disperse phase to form a curved surface is greatly enhanced. The diameter of the resulting colloidal spheres lies in the range of 10 – 80 nm, which is small enough to penetrate the narrow pore space of stones.

Frequently, ammonium acetate functionalized polysiloxanes are used as surfactants. During the drying process in the stone, small amounts of acetic acid are formed which act as a catalyst for the condensation reaction. Moreover, these functional groups can act as adhesive coupling groups to the substrate. The shorter chained cosurfactants normally consist of silanes or oligomeric siloxanes. The disperse phase is a mixture of silanes, siloxanes, and polysiloxanes.

Since microemulsions consist of reactive components that eventually form silicon resins, they do not exhibit long-term stability. However, when microemulsions are prepared with only small amounts of water, the concentration of micellar aggregates which can form ordered spatial structures is increased. Within these so-called mesophases, the water can be regarded as being dissolved in the surfactant phase. Mesophases show a markedly higher stability. Therefore, microemulsions are only sold as concentrates which are diluted with water before application. Normally, dilutions of 5 to 15 % concentrate in water are used for a water-repellent treatment of stone. Once water is added, the product should be used within 24 h.

3.2.4. Effect of Hydrophobation on Stone Properties

The range of the capillary water absorption coefficient of sandstone is on the order of $0.8\,\mathrm{kg\,m^{-2}\,h^{-0.5}}$ up to more than $20\,\mathrm{kg\,m^{-2}\,h^{-0.5}}$. After water-repellent treatment, the value decreases to $< 0.1\,\mathrm{kg\,m^{-2}\,h^{-0.5}}$, regardless of the initial value. This is equivalent to a water absorption of less than $100\,\mathrm{mL/m^2}$ during one hour of driving rain, which is a very low value.

The effect of hydrophobicity can be determined as a function of depth by storing slices cut from a drill core from a treated facade under water for 24 h (Fig. 13). To avoid sharp gradients between treated zone (dry) and the untreated material behind (moistened), a smooth increase of water absorption towards the interior should be aimed for. Whether this aim can be achieved depends on the penetration depth of the silane/siloxane solution, which in turn is influenced by the capillary absorbency of the stone.

Therefore, stones with capillary water uptake coefficient lower than $1.0\,\mathrm{kg\,m^{-2}\,h^{-0,5}}$ need no treatment with water-repellent agents because their water capacity is already sufficiently low. On the other hand, in case of a dense stone a penetration depth of the hydrophobic agent of a few millimeters has the same effect as one of several centimeters in the case of a very porous stone.

The aims of a hydrophobing treatment are:

- To prevent wet absorption of reactive gases such as sulfur dioxide and nitrous oxides
- To prevent migration and subsequent enrichment of soluble salts
- To reduce damage from hygroscopic swelling and frost
- To stop biological growth
- To diminish the amount of dust fixation and to prevent the formation of crusts.

However, in many cases, an impregnation is carried out only for esthetic reasons, i.e., to prevent re-soiling of a freshly cleaned or new facade, regardless of the absorbency and the mineralogical composition of the stone.

Laboratory investigations of hygroscopic swelling and shrinking of siloxane-treated and untreated clay-rich sandstone yielded surprising results [57]: drill-core samples of Abstwinder Schilfsandstone, fixed in a dilatometer, were stepwise wetted by rising the atmospheric humidity at 20 °C and finally stored under water for 48 h. After each step, the hygroscopic dilation and the moisture content were measured by weighing. Figure 14 shows the hygroscopic dilation of alkylsilane-treated and untreated material. The amount of water absorbed is reduced by the impregnation to some 40 % of the value for untreated stone. However, the corresponding hygroscopic dilation is enhanced for a given moisture content. In spite of a strongly reduced water uptake, the hygroscopic dilation for complete wetting is only reduced by 20 %. Under varying atmospheric humidity, treated samples react with a higher amplitude of dilation than untreated ones. A possible explanation is the following:

In the untreated material, hygroscopic dilation processes of clay minerals are transmitted into the structure to a certain extent, which depends on the pore size and geometry of the particular grains and clay matrix. If resin films are deposited in the pore space, transmission to the structure is enhanced, and the deposited material acts like a wedge and causes greater hygroscopic swelling. Since the frequency of humidity changes is much higher than that of rain fall, the risk of subsequent damage to clay-rich sandstones and limestones could be greatly enhanced.

The example shows clearly that the priority is not to exclude capillary water from the pore space but to reduce the damaging effects of water. However, a similar effect of increased hygroscopic swelling has not been measured for all organosilicon hydrophobizing products. Moreover, the same product did not behave similarly on different kinds of stone. Therefore this effect cannot be generalized, but it is recommended to test in each case whether hydrophobic treatment enhances hygroscopic swelling.

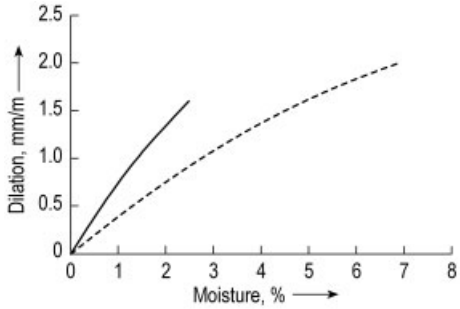

Figure 14. Hygroscopic dilation as a function of moisture absorption, determined after storing treated and untreated samples in air of different humidity and subsequent storing under liquid water to equilibrium
Full line: siloxane-treated clay-rich Schilfsandstone (Abtswind); Dotted line: untreated reference material [57].

3.2.5. Durability

Research has clearly shown that water repellents based on silicone resins show a pronounced drop in efficiency after a period of some years. Especially in the surface region, they are able to prevent the capillary uptake of water only for a short period of time [58].

The effect of UV radiation on the decrease in water repellency of silicon resins does not play an important role in practice. Investigations on formerly treated sandstone facades of the Alte Pinakothek in Munich having different exposure directions did not show any notable difference

between south and north walls; the rate of decrease was approximately of the same order.

Observations on treated buildings [59] and laboratory results [60] suggest that the main reason for the reduction in efficiency is the deposition of hydrophilic dust particles in the outermost grain layers. Thus, a thin surface layer is formed which is able to take up water again. This zone is sharply separated from the intact water-repellent zone inside. The decrease in water repellency proceeds from the surface towards the interior. In most cases, the thickness of the hydrophilic surface zone is less than 1 mm, while the interior parts still show strong water repellency. Therefore, because rain or dew does not soak into the interior of the material, the surface is more strongly exposed to strong wetting and drying than in the absence of hydrophobic treatment. In addition, the following observations have been made:

1) Expecially on limestones, the enhanced wetness burden of the surface leads to intensive biological growth. Additionally, in contrast to non carbonate sandstones, the pH range of limestones (7 – 8) also favors the growth of microorganisms. However, there is no evidence that silicone resins themselves might serve as a food resource for microorganisms.
2) Cleaned and impregnated limestone facades may look clean only for a period of some three to five years, after which a brownish or grayish crust begins to form. In some cases, the fixation of dust seems to be even more intensive than on untreated facades. A comprehensive explanation for this observation is still missing. However, it seems likely that the resoiling of facades is enhanced by persistent water films that spread within the thin hydrophilic surface layer in front of the intact hydrophobic zone in the interior.
3) Isolated water drops which run down from moldings or other protruding parts of ornaments leave ugly black tracks because they do not soak into the pores of the stone.

Salts strongly diminish the durability of hydrophobic treatments because water vapor can still diffuse into the pores. Hygroscopic salts which are imbedded in a polysiloxane network are therefore still able to dissolve and recrystallize under conditions of varying relative humidity thus destroying the polysiloxane and reducing the effectiveness.

4. Acrylates

Acrylates are very important in the conservation of building materials. The most important acrylic monomers are listed in Table 1 [61].

Polymerization at or near room temperature, as required for building conservation, can be accomplished by free radical initiation with reaction initiator pairs, among which dibenzoyl peroxide (DBP), dimethylaniline (DMA), and dibenzoyl peroxide DBP dimethyl-*p*-toluidine (DMPT) are widely used. At room temperature, the mixture of DMA or DMPT with DBP produces free radicals at a conveniently rapid rate to transform the acrylic monomers into the corresponding polymer resins, whereas DBP alone causes polymerization at a reasonable rate only above 60 °C.

Mixtures of two or more acrylic compounds may be used to produce copolymers and terpolymers having properties different from those of the homopolymers. The properties of the final products are mainly influenced by the initial ratio of the starting materials and the polymerization conditions. The glass-transition temperature T_g largely determines the suitability of a acrylic polymer for conservation purposes. At 10 °C below T_g the polymer is hard and glassy, and 10 °C above, it is flexible. The T_g of an acrylate should be below or above the temperature range to which the treated object is exposed so that no variations in mechanical properties of the polymer take place.

In contrast to several other polymers, acrylates show no or only very little yellowing. Moreover, yellowing is not caused by the acrylate compound itself but by the hardener, of which amine hardeners are most prone to change their color. For this reason, if the appropriate hardener is used, acrylates will remain without discoloration even for decades. For conservation of building materials reactive acrylate and polymer acrylate systems are used.

4.1. Reactive Acrylate Systems

Reactive systems are based on monomer or oligomeric prepolymer acrylates which are

Table 1. General formula of principal acrylic monomers $H_2C{=}CXY$ and as polymer units $(–H_2C–CXY–)_n$

Structural Unit	Abbreviation	X	Y	T_g of Polymer, °C
Methyl methacrylate	MMA	CH_3	$–C(=O)O–CH_3$	ca. 100
Ethyl methacrylate	EMA	CH_3	$–C(=O)O–CH_2–CH_3$	65
n-Propyl methacrylate	nPMA	CH_3	$–C(=O)O–CH_2–CH_2–CH_3$	35
Methyl acrylate	MA	H	$–C(=O)O–CH_3$	10
Ethyl acrylate	EA	H	$–C(=O)O–CH_2–CH_3$	− 24
Acrylic acid	AA	H	$–C(=O)OH$	

polymerized within the material to be consolidated. Polymerization is catalyzed by redox pairs which are effective at room temperature. Oxidizing compounds include peroxides, chlorates, bromates, hypochlorites; reducing agents are sulfites, hyposulfites, thiosulfates, amines, mercaptans, and others [62].

In contrast to polymeric systems, reactive acrylates were never in wide spread use because in practice there are problems with adapting them to the various conservation problems. Their main application is the filling of small cracks in tiles, paving slabs, and dense limestones such as red nodulus limestones. However, because the course of polymerization is dependent on the amount of acrylate to be polymerized, the dosage of the catalyst is very difficult to adjust. If the amount of arylate to be polymerized is big, the amount of catalyst is small, and vice versa. Therefore, if for a given volume of acrylate the content of catalyst is too high, the system starts to boil and bubbles are formed which then spread over the surface. If the amount of catalyst is too low, the system reacts very slowly or polymerization even ceases, so that the solution will not solidify. In practice the problem of finding the right catalyst content arises because the width and depth of cracks and therefore the volume of acrylate varies. Furthermore, reactive acrylate systems are sensitive to moisture and oxygen, which inhibit polymerization. Because these can not be excluded during practical work, the reactivity can not be controlled with the necessary accuracy.

Moreover, all reactive systems show a large shrinkage (10 – 15 vol %) during polymerization, so that the acrylate polymer tends to detach from the crack flanks. Consequently, the gluing effect diminishes, and the secondary pores which are formed allow the further penetration of water and acidic pollutants. To avoid such drawbacks several reactive acrylates contain a small amount of acrylic acid, which acts as a surfactant and can be incorporated into the polymer network. Acrylic acid is mainly effective for limestones and mortars because the acid can affect the calcite surface and form calcium acrylate.

In the Aachen Concept of stone conservation [63–65] efforts were made to develop acrylic prepolymers alone or in combination with epoxides that do not exhibit the above mentioned disadvantages [66–68]. However, in spite of promising laboratory test results [65], [67], [69], products have not yet appeared on the market.

4.2. Polymer Acrylate Systems

Polymer acrylates are used in the conservation of stones, brick, terracotta, tiles, porcelain, wall paintings, renderings, metals such as copper, bronze and brass, and in some cases even textiles. Hardly an area of conservation can be found where acrylates have not been used to repair or stabilize fragile objects.

The best known product is Paraloid B 72 from Rohm & Haas, a ethyl methacrylate-methyl acrylate copolymer. Many similar products with different glass-transition temperatures and mechanical properties are in use. Another widely used product is Motema Finish X from Interacryl, a butyl methacrylate – methyl methacrylate copolymer which has a lower T_g (65 °C) and a lower modulus of elasticity than poly(methyl methacrylate) ($T_g = 100$ °C) [70].

Polymer acrylates are mainly applied as 2 – 10 % solutions in organic solvents such as acetone, butyl acetate, and toluene. The volatility of the solvent decisively determines the area and method of application. If greater penetration into pores or cracks is required the solvent should evaporate slowly. The effectiveness is based on the formation of a thin polyacrylate film which covers the surfaces of the treated

materials. However, polyacrylates cannot deeply penetrate into narrow-pore materials; they are retained at the surface or within the first few millimeters and tend to form a thin surface coating which is at risk of spalling off if the modulus of elasticity of the impregnated surface layer is much higher than that of the underlying material (stone, brick, rendering). As a consequence, polyacrylates are mainly used for the conservation of limestones or marbles. Especially in Italy, many objects have been treated during the past 25 years [71]. Polyacrylates can be recommended if the decayed layer that has to be consolidated is only of a few millimeter thickness and if the underlying material is unweathered and of good strength. If, however, penetration of more than a few millimeters is required, as in the case of many weathered sandstones, application of polyacrylates is dangerous because the treated surface layer is at risk of spalling off within a short time.

Several attempts have been made to combine the consolidating effect of acrylates with a hydrophobic effect. Besides the numerous private recipes that circulate among restorers, two formulations using solutions of DRI FILM (prepolymer of methyltrimethoxysilane, MTMOS, General Electric) and Paraloid B 72 became popular [72], [73]. The product is also best known under the name Bologna Cocktail. However, in each case it must be decided whether a hydrophobic effect is desired or not.

4.3. Total Impregnation with PMMA

For the time being, the technical process of total impregnation with PMMA is only be carried out by Imchemie Stone Conservation which is based in Schesslitz near Bamberg, Bavaria. The objects must be transported to the firm because the impregnation can only be executed in a special vacuum – pressure chamber.

Before the process itself, the object is dried in an oven at 60°C for several days or even weeks depending on its size. A description of the whole process can be found in [74], [75]. After cooling to room temperature, the object is placed in the vacuum-pressure chamber. The volume of the chamber is such that objects up to a size of $4 \times 2.5 \times 2$ m can be treated. The chamber is then evacuated for several days until technical vacuum is reached and the pores are emptied of air. Then the monomer, a mixture of methyl methacrylate (MMA) and a small amount of butyl methacrylate, is let into the chamber until the object is totally immersed. The MMA fills all pores and penetrates to the very core of the stone. Then the chamber is flooded with nitrogen and the pressure is raised to 10 bar, so that excess MMA is pumped back into the reservoir. The polymerization is started by raising the temperature.

The process of polymerization is controlled by a variety of peroxides which decompose at certain temperatures (the exact composition of the peroxides is a trade secret). The polymerization is a strongly exothermic process which under normal conditions would lead to boiling of the liquid. Therefore, the treatment is carried out under pressure so that the formation of gas bubbles is suppressed.

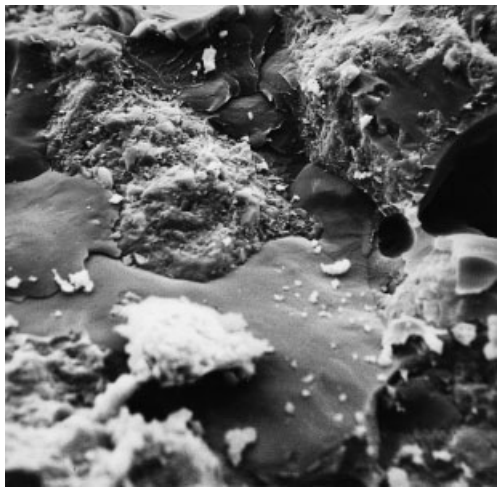

Figure 15. Ansbacher Schilfsandstone, treated with the Ibach Method. The PMMA completely fills the entire pore space. Isolated shrinkage bubbles can be observed within the PMMA.

At the end of the polymerization process, the temperature has reached ca. 70 °C. Polymerization normally takes one week or even longer. When polymerization is complete, the object is allowed to cool slowly to room temperature over several days. After returning to normal pressure, the object can be removed from the chamber. The whole pore space is filled with solid PMMA, and the formerly porous material no longer absorbs water. Figure 15 shows a SEM

image which demonstrates the complete pore filling with PMMA in the case of a treated sandstone. Within the PMMA, isolated vacuoles can be recognized whose formation is caused by shrinkage in the course of polymerization. The shrinkage bubbles only occur within the PMMA and not at the interface between PMMA and pore wall, because the coupling agents which are added to the impregnation solution prevent the detachment of the PMMA from the pore walls. Even experts are unable to distinguish treated objects from untreated ones by mere visual inspection because their appearance remains unchanged.

Since 1978 total impregnation with PMMA has been frequently applied in Bavaria for sandstone objects, such as tombstones, balusters, garden benches, and sculptures. By general agreement among restorers and public conservation authorities, important works of art are still excluded from treatment with the Ibach method because it is irreversible. Only objects whose condition is so poor that conservation with conservative means will not provide the necessary durability are treated.

The Ibach method has now been applied to marble in the case of the niche and gable figures of the Glytothek in Munich. Several investigations were carried out to transfer the process from sandstone to marble consolidation, because appropriate coupling agents had to be found. Coupling agents control the shrinkage of 19 vol % which occurs during the polymerization process. The coupling agents ensure that the shrinkage bubbles do not occur between the mineral surface and the PMMA, but only as isolated vacuoles within the PMMA itself if the correct concentration of coupling agents is added [76]. Only after a sample treatment with a figure of 1 m in size had yielded a perfect penetration, was the first monumental figure of Peter Vischer from the Glytothek subjected to total impregnation in 1983.

5. Epoxy Resins

Epoxy resins are two-component plastics which consist of the epoxy resin itself and a hardener (→ Epoxy Resins). Bisphenol A resins are widely used in industry and building conservation.

Bisphenol A reacts with epichlorohydrin in two steps to form Bisphenol A diglycidyl ether.

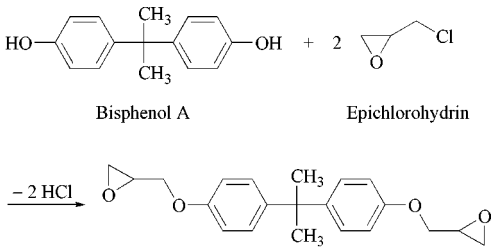

Bisphenol A diglycidyl ether

The Bisphenol A diglycidyl ether molecules are condensed into linear long-chain molecules which constitute the Standard Bisphenol A epoxy resins (see bottom).

Cold-hardening epoxy resins, which are of interest for building-conservation purposes, are transformed into solid epoxy resin by means of tetrafunctional aliphatic amine hardeners. The addition of the amine to the epoxy group leads to β-hydroxyamine links between resin and hardener and the formation of a spatial network (Fig. 16).

To adapt epoxides to the requirements of stone conservation, several experiments have been conducted within the framework of the Aachen Concept of stone conservation [63], [64], [66], [67]. Among them, a modification of the epoxide system with an glycidylmethacrylate (GMA), a methacrylate with an epoxy side chain, yielded promising results, because the hardening rate as well as the mechanical properties of the resulting resin could be successfully influenced [68]. More detailed information about the possibilities of the molecular design

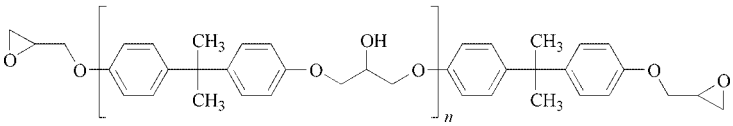

Standard Bisphenol-A epoxy resin

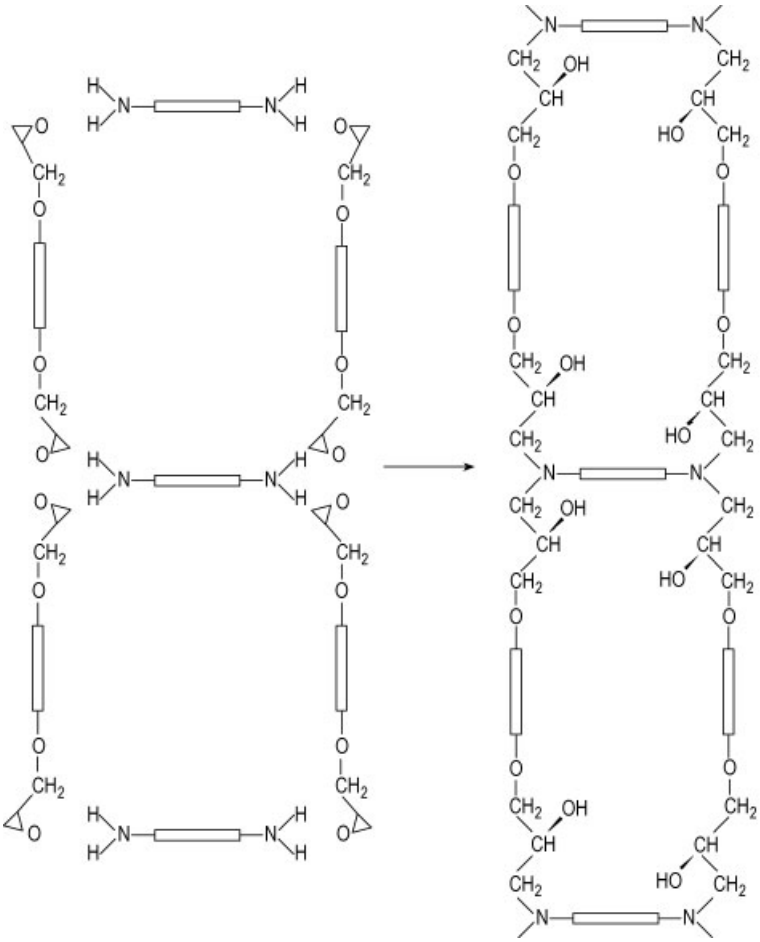

Figure 16. Epoxy resin formation with tetrafunctional amines

of epoxides can be found in [77–79]. The laboratory experiments were accompanied by SEM studies on impregnated stones in order to assess the quality of the polymer films within the pore space [80].

In building conservation, epoxy resins are used because of their high adhesive strength. Depending on the type of epoxy resin, an adhesive strength of up to 30 MPa or even more can be obtained. For this reason, the main field of application is the repair of broken pieces that must be glued together or the fixation of dowels or anchors.

Among scientists and restorers, contradicting opinions still exist about the suitability of epoxy resins for consolidating decayed stones. In particular in the former German Democratic Republic, several natural stone outdoor sculptures were consolidated with epoxy resins, which still seem effective after almost 20 years of exposure. However, because of the insufficient penetration depth of most epoxides, treatment was carried out by means of pressure injection, for which many holes had to be drilled to allow injection of the resin deep into the stone.

One of the major risks in the use of regular commercial epoxides for consolidation of decayed stones is the separation of the resin and the hardener due to chromatographic effects within the narrow pore space of the stone, whereby one of the components can be retained at the pore surfaces while the other component penetrates deeper [81]. As a consequence, the resin will not or only insufficiently harden without having

an consolidating effect. A further drawback of normal epoxides is their proneness to yellowing, which can cause ugly discolorations. Moreover, the modulus of elasticity of regular commercial epoxy resins, which were designed for other purposes, is too high, so that consolidated surface parts of stones are at risk of spalling off with time.

For this reason, consolidation of stone with epoxy resins is still in the research stage [63–65], [67], [79]. The selective adsorption of the resin or hardener can be prevented by preparing a quasi-one-component system by controlled linking the two reaction partners. Using a tetrafunctional hardener with a bifunctional resin (Fig. 16) yields highly linked macromolecules which result in high viscosity of the liquid. Sufficient penetration into a pore system is no longer possible. To produce defined quasi-one-component epoxy systems, the resin or the hardener must be chemically modified so that only two resin molecules are connected to one hardener molecule. This can, for instance, be achieved by replacing one of the epoxy groups with an activated Michael system, which is more reactive than the epoxy group [79].

Moreover, by inserting a further Bisphenol F unit into the resin molecular chain a lowering of the modulus of elasticity of the final product is achieved (see below).

The effect of epoxy resins on the mechanical properties of selected natural stones in comparison to other consolidants has been extensively investigated [82].

Epoxy resins are suitable for preparing stone-repair materials. Because the mixing ratio of aggregate and binder can be lowered to 1 : 15, the appearance of the resulting mortar is very similar to that of stone. The advantage of epoxy-bound repair mortars is that their color can be more easily adapted to a natural stonelike appearance than that of lime or cement-bound mortars because the epoxy resin is colorless and its proportion in the mixture is lower than in the case of inorganic binding materials. Mortars with a low binder to aggregate ratio show a high porosity; their capillarity, which is similar to that of natural stones, strength, and modulus of elasticity are moderate. However, repair mortars based on epoxy resins also exhibit the problem of yellowing under outdoor exposure, and this can lead to an ugly patchy appearance of treated facades.

Epoxy-based mortars with a higher resin content are very dense and can be polished. Therefore they can be used for repairing decorative stones like colored limestones, schists or marbles. In these cases the result is mostly much better than with lime- or cement-based mortars.

6. Polyurethanes

Isocyanates and polyalcohols are the starting components for producing polyurethane resins. The reaction product is a polyurethane chain

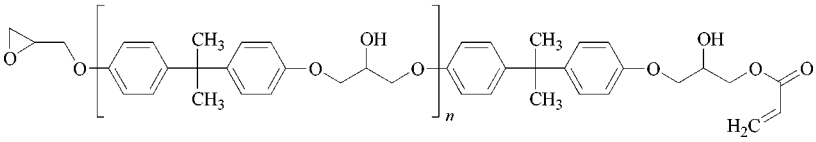

Activated Bisphenol A resin

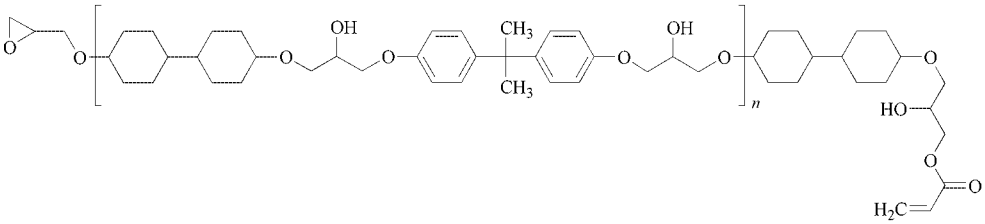

Epoxy group Hydrogenated Bisphenol F Bisphenol A Reactive Michael system

Activated, flexible Bisphenol resin

which can further polymerized to a polyurethane network.

To adapt polyurethanes to the requirements of stone conservation, intensive research has been undertaken within the joint research program "Steinzerfall – Steinkonservierung" (German: stone decay/stone conservation), funded by the German Federal Ministry of Education and Research (BMBF). The research, which was largely focused on the development of modified polyurethanes became known as the Aachen Concept for stone conservation [63], [64], [66], [67]. Because monomer solutions exhibit several problems in case of stone conservation, prepolymers must be used which consist of hard and soft segments to lower the elasticity of the resulting film (Fig. 17).

In the first step, a segmented prepolymer is formed by the reaction of a diisocyanate and a diol. The elastic properties are mainly determined by the length and the chemical nature of the soft segment, which in most cases is located in the diol component. In the Aachen Concept [79], the segmented prepolymer consists of isophorone diisocyanate as the hard segment and a hydroxy-functionalized oligomethylsiloxane as the soft segment (silane-modified polyurethane).

Depending on the desired reactivity, the prepolymers can contain from 0.1 to almost 10 wt % of monomer. The prepolymers are dissolved in organic solvents, which also determine properties such as penetration and reactivity. The amount of polyurethane compound in the product is on the order of 30 wt %.

The segmented prepolymer is applied to the stone, and the reactive O=C=N – groups at both ends of the molecule react with the humidity within the pore space to form the final polyurethane film. A byproduct of the reaction is CO_2, which evaporates. Due to the methylsiloxane component, the film has hydrophobic properties.

By varying the length of the soft-segment chain, the mechanical properties of the polyurethane film can be deliberately controlled. For a given length of the hard segment, polymer films with a $C_3 - C_4$ soft segment are brittle, and those with $C_9 - C_{12}$ segments are fluid. Only those with C_5 or C_6 soft segments are flexible and appropriate for stone conservation.

The film formation of polyurethanes in the presence of different salts has been investigated [80], [83]. The reaction rate is lowered, and the films no longer cover all salt crystal surfaces completely, so that salts can be mobilized by moisture. For this reason, in the presence of salts the use of a catalyst such as dibutyltin dilaurate is necessary. The influence of Aachen-type polyurethanes on the mechanical and moisture-transport properties of impregnated stones was investigated [65], [82], and the effect of polyurethane treatment on the swelling and shrinking of natural stones under different moisture conditions measured [84]. Information about the durability of treatments in time-lapse laboratory tests can be obtained from [85].

7. Polyester and Alkyd Resins

Unsaturated polyesters (UP) resins consist of a multibasic, unsaturated carbon acid (phthalic acid, maleic acid, fumaric acid, etc.) and dihydric alcohol, such as ethylene glycol, propanediol, or butanediol (→ Polyester Resins, Unsaturated). The two components are dissolved in monomers which form bridges between the double bounds of the acid and the alcohol. These include styrene, methylstyrene, and methyl methacrylate. The reaction is catalyzed by peroxides, Co salts, or tertiary amines.

Alkyd resins (→ Alkyd Resins) basically consist of the same or similar unsaturated acids and alcohols, which are dissolved in reactive substances such as styrene, acrylates, or epoxides that can link the polyester chains into a three-dimensional network. Alkyd resins are modified with 15 – 75 % fatty acids or oils such as linseed oil, soybean oil, or castor oil.

In stone and building conservation, polyesters are mainly used as constituents in paints (alkyd paints) or as binders in stone-repair mortars. Especially in the 1960s and 1970s, many repairs on limestones and marbles were carried out with polyester-bound materials. However, polyester repair mortars did not prove as durable as expected because their thermal expansion is much greater than that of all mineral building materials, so that sooner or later the repair patches detached from the substrate. Moreover, the repair mortars become brittle and bleach out because of ongoing cross-linking of the chains

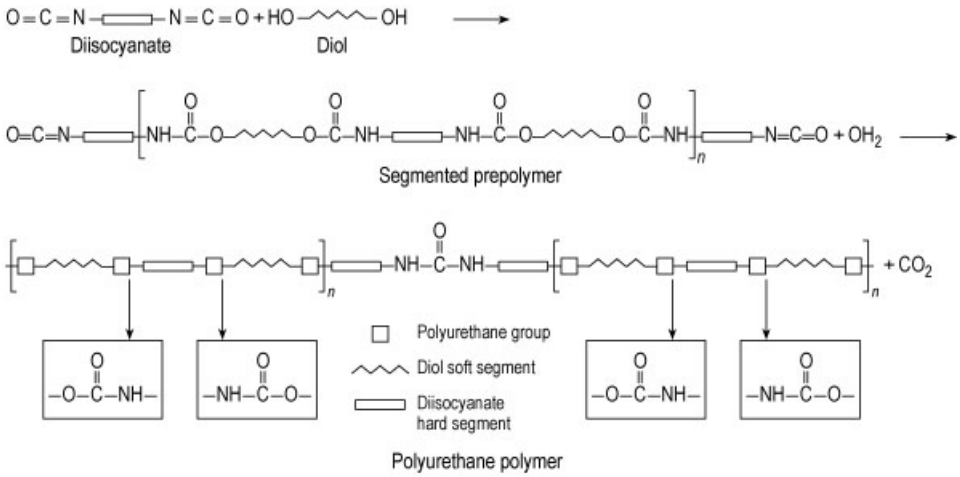

Figure 17. Reaction path for produce a polyurethane stone-conservation product

and oxidation of hydrocarbon molecules. Today, the majority of polyester repairs have already fallen out or are so loose that they can be removed very easily. In restoration and conservation, polyester repairs should be removed and replaced by mineral-based repair mortars.

Polyester repair mortars are still used for filling defects in pavement or facade slabs made of natural stone. Filling of the defects which occur in many stones is often necessary for polishing. If the polyester repairs are carried out appropriately they can hardly be recognized after polishing. However, the method can not be recommended for large facades where the slabs are exposed to strong light, wind, and rain because the repairs would detach in a quite short time. Polyesters are also very used for gluing steel anchors by which facade plates are fixed.

Between 1960 and 1970 in Germany, the polyester product Lemisan was used as consolidant for weathered sandstones and limestones. Lemisan was an unsaturated polyester dissolved in styrene and diluted to a concentration of 10 – 15 % with dichloromethane. The product was not a special development but corresponded to the product Leguval W 16 from Degussa, which was simply diluted to adapt its properties to the needs of stone conservation (capillary penetration into porous materials). Because Lemisan did not penetrate deeply enough into the stones, in several cases the detachment of scales parallel to the surface was observed whose thickness exactly reflects the penetrated and consolidated zone. An example is shown in Figure 18.

Figure 18. Thin desquamations on a polyester-treated sandstone facade. Parish church in Oberdachstätten near Ansbach, Bavaria.

8. Polymer Dispersions

In building conservations, several aqueous polymer dispersions are used to modify the properties of inorganic materials such as mortars or lime paints and to adapt them to special requirements. For instance, polymers like poly(vinyl acetate), acrylates, polyurethanes, and epoxides are also available in aqueous dispersion. However, this method of improving the workability of a well-known inorganic systems is not an invention of the 1990s. Already in antiquity, craftsmen used natural polymers such as casein, collagen, egg albumin, and blood albumin, whose ef-

Table 2. Properties of polymer dispersions which are most frequently used in building conservation

Polymer ∗	Type	Particle size, µm	Solid content, wt %	Viscosity, mPa · s	Density, g/cm^3	T_c film formation ∗∗	pH value
PVP copolymer	nonionic	0.5 – 3	50	1500 – 3500	1.08 – 1.09	20 – 30	4 – 6
PVA copolymer	nonionic	0.1 – 2	50	500 – 3000	1.07 – 1.2	10 – 30	4 – 7
PVDC copolymer	anionic	0.1	55	10 – 20	1.26	4	2
Acrylate – acrylnitrile copolymer	anionic	0.3	50	100 – 220	1.01	1	4 – 6
Acrylate styrene copolymer	anionic	0.1	50 – 60	200 – 2000	1.03	1 – 2	7 – 9
Acrylates	anionic	0.1 – 0.6	50 – 55	10 – 1500	1.01 – 1.09	1 – 30	5 – 9
PU	nonionic	0.1 – 3	35 – 50	500 – 3000	1.0 – 1.1	10	6 – 7
EP	nonionic	0.1 – 5	40 – 60	500 – 3000	1.06 – 1.15	10	6 – 7

∗ PVP = poly(vinyl propionate), PVA = poly(vinyl acetate), PVDC = poly(vinylidene chloride), PU = polyurethane, EP = epoxy resin.
∗∗ Critical temperature of film formation.

fects on different mortars were intensively investigated [86]. Table 2 presents an overview over the most common modern polymer dispersions.

Because polymer dispersions can be diluted with water to any desired degree, their properties can vary over a very broad range. Moreover, polyurethane or acrylic dispersions, for instance, are large classes of chemical substances whose properties differ greatly, so that the figures presented in Table 2 is only a rough guide of the respective properties.

To synthesize polymer dispersion in an aqueous medium, special preparation techniques are necessary. Prepolymers are dispersed in water as small droplets and surrounded by an protective colloid. Specific catalysts within the prepolymer droplets then start the polymerization to form a solid polymer particle. The stability of the resulting emulsion is secured by the presence of protective colloids which surround the polymer particles, whose size is in the range of 0.1 to 5 µm. Because the size of the polymer particles is in the order of the wavelength of visible light, the emulsions are cloudy of milky white due to the scattering of light.

The major advantage of polymer dispersions is that they can be added to all classical building materials which are worked with water. When applied to a building material, the dispersion breaks in the course of evaporation of water, forming a polymer film which covers the surface of the mineralic substrate.

The main fields of application are as follows:

- Paints: dispersion lime or silicate paints for indoor or outdoor application
- Mortars: polymer aided mortars for various applications such as repair mortars, joint mortars, spatala materials, fixation of wall and floor tiles
- Renderings: polymer-aided renderings
- Concrete: polymer-aided concrete where particular requirements with respect to damp-proofness, acidic resistance, and solvent resistance exist
- Primers: primer layer of polymer dispersion to improve the adhesion of lime or hydraulic mortars to mineral building materials.

The effect on the mechanical properties of the mortars is significant. Whereas the modulus of elasticity is lowered in most cases, bending strength as well as compressive strength increase markedly. However, there is no linear increase in bending and compressive strength with increasing polymer content. The most striking influence is observed in case of adhesive strength (pull-off strength) which can reach values of 1 – 2 MPa or more. It increases almost linearly with increasing polymer content.

Modern stone-repair mortars based on polymer-aided cement as binder show a modulus of elasticity as low as that of lime mortar (4 – 6 kN/mm^2) and a strain that is about ten times higher than that of regular mineral mortars ($\varepsilon = 0.2 - 0.5$ mm/m). Tables 3 and 4 contain some illustrative data from [87], [88]. The results clearly demonstrate that the effect of polymer additives on different mortars cannot be simply predicted but have to be measured in each case.

However, the disadvantages of high polymer content in building materials must also be mentioned. The polymer film shows much greater

Table 3. Modulus of elasticity, compressive strength and bending strength of polymer-modified hydraulic lime mortars [87], [88]

Mortar/additive [a]	Additive content, wt %	E_{dyn}, MPa [b]	β_{BS}, MPa [c]	β_{CS}, MPa [d]
HL/-	0	10900	1.9	9.5
HL/DP 1	3	11700	3.0	9.7
HL/DP 1	5	9900	2.7	8.6
HL/DP 1	7	9200	3.0	8.8
HL/DP 2	3	12100	3.3	11.5
HL/DP 2	5	11000	3.3	10.0
HL/DP 2	7	10700	3.4	9.3
HL/DL 1	3	10600	2.9	11.0
HL/DL 1	5	10500	3.2	11.6
HL/DL 1	7	11900	4.0	15.7
HL/DL 2	3	9300	2.1	8.3
HL/DL 2	5	8100	1.9	7.3
HL/DL 2	7	7900	1.9	7.5

[a] Composition of the mortar: 75 % hydraulic lime (water/lime hydrate) + 25 % blast furnace slag as pozzolanic additive.
DP 1: dispersion powder copolymer of vinyl acetate and ethylene with mineral additives; protective colloid poly(vinyl alcohol).
DP 2: dispersion powder terpolymer of vinyl acetate, vinyl versatate, butyl acrylate; protective colloid poly(vinyl alcohol).
DL 1: liquid dispersion copolymer of methyl methacrylate and butyl methacrylate.
DL 2: liquid dispersion copolymer of styrene and acrylic ester.
[b] E_{dyn}: dynamic modulus of elasticity.
[c] β_{BS}: bending strength.
[d] β_{CS}: compressive strength.

Table 4. Effect of polymer dispersions on mechanical properties of cement mortars $*$

Polymer $**$	Content, wt %	β_{BS}, MPa (w/c ratio)	β_{CS}, MPa (w/c ratio)	β_{POS}, MPa (w/c ratio)
–	0	6.5 (0.45)	35 (0.45)	0.2 (0.70)
PVP 1	2.5			0.6 (0.70)
PVP 1	5	9.1 (0.45)	44 (0.45)	1.0 (0.70)
PVP 1	10	9.8 (0.45)	44 (0.45)	1.7 (0.70)
PVP 1	25	10.5 (0.45)	43 (0.45)	2.5 (0.65)
PVP 2	5	9.3 (0.45)	44 (0.45)	n.d.
PVP 2	10	10.4 (0.45)	44 (0.45)	1.5 (0.55)
PVP 2	25	11.3 (0.45)	41 (0.45)	2.2 (0.45)
PA	2.5	n.d.	n.d.	0.4 (0.70)
PA	5	n.d.	n.d.	1.2 (0.60)
PA	10	10.2 (0.45)	34 (0.45)	2.2 (0.50)
PA	25	n.d.	n.d.	2.7 (0.50)

$*$ β_{BS}: bending strength, β_{CS}: compressive strength, β_{POS}: pull-off strength, w/c ratio: water/cement ratio. $**$ PVP: poly(vinyl propionate), PA: polyacrylate

hygroscopic swelling under changing moisture conditions than mineral building materials. For this reason, paints with a high content of dispersion very often peel off when the pull-off strength of the underlying rendering is insufficient. Likewise the water-vapor permeability decreases with increasing polymer content, a particularly problematic circumstance regarding the durability of dispersion paints. If small cracks form in the paint, large amounts of water can penetrate deep into the underlying rendering. Its evaporation, however, is then hindered by the high polymer content of the paint. As a consequence, the water content at the interface between rendering and paint increases steadily, thus destroying the outermost layer of the rendering. The paint layer, which responds strongly to temperature and humidity variations, soon delaminates from the powdering and sanding surface of the rendering.

In general, polymer-modified building materials show different aging behavior to regular mineral materials. This is caused by the longer water retention and an enhanced tendency for soiling, which is in particular exhibited by building materials with high polymer content. The problem of long-term compatibility of polymer-modified and regular building materials is exemplified by natural stone buildings for which after only a few years the polymer-containing repair mortars stand out from the facade as dark patches.

In spite of the small particle size, polymer dispersion are not appropriate for consolidating decayed natural stones or renderings. The polymer particles are retained at the surface, where they form an impermeable film. Only the water penetrates deeper into the pore space of the respective material. The polymer film lends the surface an undesired shiny appearance and prevents water-vapor transport. Moreover, the strength and the modulus of elasticity of the impregnated surface layer increase so that it is under risk to spall off.

9. Conclusions

Chemical conservation of buildings is a relatively new field of research. For this reason, long-term experience with the durability of polymers in contact with mineral materials is not available. There is no doubt that polymer additives can improve the workability of traditional materials and enlarge the area of their application. Likewise, chemicals can restore decayed stones and protect them from rain and frost. However, chemical building conservation is not just a theoretical science, but is strongly dependent on feedback from practitioners, and the rules of correct craftsmanship must always be respected. The feedback from practitioners helps to define aims of chemical research. Respecting the rules of good craftsmanship will help to avoid building mistakes which can not be repaired by chemical means. With respect to these considerations, refs. [89–91] can be recommended.

10. References

1. E. M. Winkler: *Stone, Properties, Durability in Man's Environment,* Springer Verlag, Heidelberg – New York – Wien 1975.
2. P. Brimblecombe, *Eur. Cult. Herit. Newsl. Res.* **2** (1988) no. 2, 13 – 14.
3. S. Luckat: Quantitative Untersuchungen des Einflusses von Luftverunreinigungen bei der Zerstörung von Naturstein, Forschungsbericht 104 01 018/02, Umweltbundesamt Berlin, 1982, p. 139.
4. F. Giradet, V. Furlan, *Proc. Int. Congr. Deterior. Conserv. Stone, Proc. 4th* (1982) 127 – 134.
5. T. P. Cooper, Proceedings European Symposium Science, Technology and European Cultural Heritage, Bologna 1989, pp. 465 – 468.
6. P. A. Baedecker, E. O. Edney, P. J. Moran, T. C. Simpson, R. S. Williams: Effects of Acidic Deposition on Materials, Sate-of-Science/Technology Report 19, National Acid Precipitation Assessment Program NPAP, Washington 1990.
7. S. Brüggerhoff, P. W. Mirwald, *Proc. Int. Congr. Deterior. Conserv. Stone 7th* (1992) 715 – 724.
8. V. Furlan, F. Giradet, UNESCO RILEM Symposium Conservation of Stone and Other Materials Paris, E & FN Spon, London 1993, pp. 27 – 34.
9. M. M. Reddy, S. Sherwood, B. Doe, *Proc. Int. Congr. Deterior. Conserv. Stone 5th* (1985) 517.
10. N. S. Baer, Scientific Evaluation of NAPAP's Final Report, PA: Air and Waste Management Assessment, Pittsburgh 1991, pp. 153 – 164.
11. R. N. Butlin et al., *Proc. Int. Congr. Deterior. Conserv. Stone 7th* (1992) 345 – 354.
12. R. A. Livingston, *Proc. Int. Congr. Deterior. Conserv. Stone 7th* (1992) 375 – 386.
13. R. A. Livingston, *Proc. Int. Congr. Deterior. Conserv. Stone 5th* (1985) 509 – 516.
14. L. G. Johansson, O. Lindqvist, R. Mangio, *Durability of Building Materials* **5** (1988) 429 – 449.
15. B. Hicks: Conservation of Historic Stone Buildings and Monuments, Report of the Committee on the Conservation of Historic Stone Buildings and Monuments, National Academy Press, Washington D.C. 1982, pp. 183 – 196.
16. W. Dannecker: *Wechselwirkungen zwischen technischen Oberflächen und Atmosphäre,* VDI Verlag, Düsseldorf 1989, pp. 237 – 264.
17. M. Steiger et al.: *Jahresbericht Steinzerfall – Steinkonservierung* **4,** Verlag Ernst & Sohn Berlin 1992, pp. 215 – 239.
18. M. Steiger, F. Wolf, W. Dannecker, UNESCO RILEM Symposium Conservation of Stone and Other Materials, vol. 1, E & FN Spon, London 1993, pp. 35 – 42.
19. C. Wittenburg, W. Dannecker, *J. Aerosol Sci.* **23** (1992) Suppl. 1, 869 – 872.
20. E. Bock, P. G. Josza, W. Sand, R. Mansch, M. Wilimzig, *Denkmalpflege und Naturwissenschaft, Natursteinkonservierung,* **II,** Fraunhofer IRB Verlag, Stuttgart 1998, pp. 137 – 172.

21. W. E. Krumbein: *Jahresberichte Steinzerfall – Steinkonservierung,* **3,** Verlag Ernst & Sohn, Berlin 1991, pp. 215 – 229.
22. W. E. Krumbein, *Denkmalpflege und Naturwissenschaft, Natursteinkonservierung,* **II,** Fraunhofer IRB Verlag, Stuttgart 1998, pp. 173 – 205.
23. R. A. Livingston: *Saving Our Architectural Heritage – The Conservation of Historic Stone Structures,* Dahlem Workshop Report, John Wiley & Sons, Chichester – New York 1997, pp. 37 – 62.
24. P. W. Mirwald, S. Brüggerhoff: *Saving our Architectural Heritage – The Consewrvation of Historic Stone Structures,* Dahlem Workshop Report, John Wiley & Sons, Chichester – New York 1997, pp. 255 – 268.
25. F. Rathgen: *Die Konservierung von Altertumsfunden,* **vol. 1: Stein und steinartige Stoffe,** Spemann, Berlin 1915.
26. S. Z. Lewin, *Art Archaeol. Tech. Abstr.* **6** (1966) no. 1, 185 – 277.
27. J. Riederer, D. Beeger, E. Kühne, Berliner Beiträge zur Archäometrie, **12,** Staatliche Museen zu Berlin, Rathgen Forschungslabor, Berlin 1993, pp. 115 – 202.
28. R. J. Schaffer: "The Weathering of Natural Building Stones," Building Research Establishment spec. report **18,** Garston-Watford (1931, reprint 1972).
29. C. Herm, S. Pfefferkorn, R. Snethlage, *Denkmalpflege und Naturwissenschaft, Natursteinkonservierung,* **vol. II,** Fraunhofer IRB Verlag, Stuutgart 1998, pp. 9 – 26.
30. M. Koller, J. Nimmrichter, H. Paschinger, H. Richard, *Restauratorenblätter* **17** (1996) 143 – 156.
31. J. M. Teutonico, I. McCaig, C. Burns, J. Ashurst, 2nd Intern. EUROLIME meeting, EUROCARE Secretariat Uni Karlsruhe 1993, pp. 71 – 77.
32. A. Wolff: "35. Dombaubericht," *Kölner Domblatt* **59** (1994) 16; "36. Dombaubericht," *Kölner Domblatt* **60** (1995) 16; "37. Dombaubericht," *Kölner Domblatt* **61** (1996) 16.
33. A. Stois, *Veröffentl. Ges. Bayer. Landeskd.* **7 – 12** (1935) 199 – 224.
34. A. W. Keim: *Die Mineral-Malerei. Neues Verfahren zur Herstellung witterungsbeständiger Wandgemälde,* Technisch-Wissenschaftliche Anleitung, Wien 1881, Reprint Thomas Hoppe, Stuttgart 1995.
35. J. Osswald, *ID Veröffentlichungen des Instituts für Denkmalpflege ETH Zürich* **19** (1998) 147 – 156.
36. H. Ettl, L. Sattler, H. Schuh: *Denkmalpflege und Naturwissenschaft, Natursteinkonservierung,* **vol. I,** Verlag Ernst & Sohn, Berlin 1996, pp. 105 – 126.
37. C. A. Grissom, N. R. Weiss, *IIC Technical Abstracts* **18** (1981) no. 1, 150 – 220.
38. M. Wick, G. Kreis, F. H. Kreuzer, *Ullmann's,* 4th ed., **vol. 21,** pp. 485 – 543.
39. W. Beier, U. Wellbrock, G. H. Frischat, *Glastechnische Berichte* **63** (1990) no. 6, 168.
40. L. Sattler, Thesis Universität München, published by Bayerisches Landesamt für Denkmalpflege Zentrallabor, Forschungsbericht 9, München 1992, p. 156.
41. R. K. Iler: *The Chemistry of Silica,* John Wiley & Sons, New York 1979, p. 866.
42. W. Beier, M. Meier, G. H. Frischat, *Glastechn. Ber.* **58** (1985) no. 5, 98.
43. G. Grassegger, *Jahresberichte Steinzerfall – Steinkonservierung,* **2,** Verlag Ernst & Sohn, Berlin 1992, pp. 179 – 185.
44. G. Grassegger, S. Adam, *Jahresberichte Steinzerfall – Steinkonservierung,* **4,** Verlag Ernst & Sohn, Berlin 1992, 127 – 133.
45. E. Wendler: *Saving our Architectural Heritage, The Conservation of Historic Stone Structures, Report of the Dahlem Workshop,* John Wiley & Sons, New York 1997, pp. 181 – 198.
46. R. Snethlage: Steinkonservierung 1979 – 1983, Arbeitshefte des Bayerischen Landesamtes für Denkmalpflege 22, München 1984, p. 203.
47. E. S. Goins, G. S. Wheeler, M. T. Wypyski, *Proc. Int. Congr. Deterior. Conserv. Stone 8th* (1996) 1255 – 1264.
48. E. Wendler, D. D. Klemm, R. Snethlage, *Proc. 5th Int. Conf. Durability of Building Materials and Components*, Chapman & Hall, Brighton 1991, pp. 203 – 212.
49. C. H. Chiang, H. Ishida, J. L. Koenig, *J. Coll. Interface Sci.* **74** (1980) no. 2, 396 – 404.
50. L. Sattler, R. Snethlage, *Denkmalpflege und Naturwissenschaft, Natursteinkonservierung,* **vol. I,** Verlag Ernst & Sohn, Berlin 1996, pp. 89 – 104.
51. E. Wendler, R. Rückert-Thümling, 3. Int. Koll. Werkstoffwissenschaften und Bausanierung, Technische Akademie Esslingen 1992, pp. 1818 – 1830.
52. L. Sattler, *Jahresberichte Steinzerfall – Steinkonservierung,* **2,** Verlag Ernst & Sohn, Berlin 1992, pp. 165 – 168.

53. M. Boos, J. Grobe, E. Hilbert, E. Wendler, *Bautenschutz & Bausanierung* **8** (1997) 39 – 41.
54. M. Roth, *Kunststoffe im Bau* **19** (1984) no. 2, 71 – 74.
55. K. L. Mittal: *Silanes and Other Coupling Agents,* VSP Intern. Science Publ., Zeist 1992, p. 92.
56. H. Mayer, M. Roth, *Bausanierung* **13** (1990) 1 – 4.
57. U. Möller, H. Schuh, E. Wendler, 3. Int. Koll. Werkstoffwissenschaften und Bausanierung, Technische Akademie Esslingen 1992, pp. 1796 – 1806.
58. E. Wendler, Proc. 1st Int. Symp. Surface Treatment of Building Materials with Water Repellent Agents, Delft 1995, pp. 25-1, 26-14.
59. E. Wendler, R. Snethlage, Proc. Int. Symp. The Engineering Geology of Ancient Works, Monuments and Historical Sites, vol. 2, Balkemaa, Rotterdam 1988, pp. 945 – 951.
60. E. De Witte, H. De Clercq, R. De Bruyn, A. Pien, Proc. 1st Int. Symp. Surface Treatment of Building Materials with Water Repellent Agents, Delft 1995, pp. 5-1, 5-10.
61. E. W. Merrill, *Mater. Res. Soc. Symp. Proc.* **462** (1997) 353 – 358.
62. H. Rauch-Puntigam, T. Völker: *Acryl- und Methacrylverbindungen. Chemie, Physik und Technologie der Kunststoffe in Einzeldarstellungen,* **vol. 9,** Springer-Verlag, Berlin – Heidelberg – New York 1967.
63. D. Honsinger: "Strukturmerkmale polymerimprägnierter Sandsteine," Thesis, Fakultät Bauingenieurwesen und Vermessungstechnik, RWTH Aachen 1990, p. 218.
64. D. Honsinger et al., *Jahresberichte Steinzerfall – Steinkonservierung* **2,** Verlag Ernst & Sohn, Berlin 1992, pp. 153 – 163.
65. K. Littmann, B. Riecken, H. R. Sasse, *Denkmalpflege und Naturwissenschaft, Natursteinkonservierung,* **vol. II,** Fraunhofer IRB Verlag, Stuttgart 1998, pp. 207 – 226.
66. S. Wagener, A. Hessland, H. Höcker, *Jahresberichte Steinzerfall – Steinkonservierung* **2,** Verlag Ernst & Sohn, Berlin 1992, pp. 139 – 151.
67. K. Littmann, G. Pleyers, B. Riecken, H. R. Sasse, B. Schwamborn, *Jahresberichte Steinzerfall – Steinkonservierung* **4,** Verlag Ernst & Sohn, Berlin 1994, pp. 87 – 101.
68. U. Schneider, S. Wagener, H. Höcker, *Jahresberichte Steinzerfall – Steinkonservierung* **4,** Verlag Ernst & Sohn, Berlin 1994, pp. 117 – 125.
69. K. Littmann, B. Riecken, H. R. Sasse, B. Schwamborn, *Jahresberichte Steinzerfall – Steinkonservierung* **3,** Verlag Ernst & Sohn, Berlin 1991, pp. 85 – 94.
70. P. J. Koblischek, Proceedings 9th Workshop EUROCARE-EUROMARBLE, Forschungsberichte Bayerisches Landesamt für Denkmalpflege Zentrallabor 17/1999, München 1999, pp. 109 – 120.
71. G. Alessandrini, M. Laurenzi-Tabasso, Proceedings 6th Workshop EUROCARE-EUROMARBLE, Forschungsberichte Bayerisches Landesamt für Denkmalpflege Zentrallabor 16/1996 München, pp. 1 – 20.
72. O. Nonfarmale, *Conserv. Stone, Proc. Int. Symp.,* (1975) 401 – 410.
73. R. Rossi-Manaresi, G. Alessandrini, S. Fuzzi, R. Peruzzi, *Proc. Int. Congr. Deterior. Conserv. Stone 3rd* (1979) 367 – 376.
74. R. Snethlage, R. Wihr, Arbeitsheft 4: Steinkonservierung, Bayerisches Landesamt für Denkmalpflege, München 1979, pp. 26 – 34.
75. P. Sinner, Science, Technology and Cultural Heritage, Proceedings of the European Symposium, Bologna 1989, Butterworth-Heinemann, Oxford 1992, pp. 700 – 703.
76. H. G. Lorenz, Proceedings of the 7th Workshop EUROCARE-EUROMARBLE, ICE/HT-FORTH, Patras 1997, pp. 193 – 198.
77. H. Reul: *Handbuch der Bauchemie,* Verlag H. Ziolkowsky, Augsburg 1991, p. 370.
78. N. Hoffmann, H. Höcker, *Jahresberichte Steinzerfall – Steinkonservierung* **6,** Fraunhofer IRB Verlag, Stuttgart 1998, pp. 45 – 54.
79. G. Gerhard-Abozari, N. Hoffmann, M. Viol, H. Höcker, *Jahresberichte Steinzerfall – Steinkonservierung* **6,** Fraunhofer IRB Verlag, Stuttgart 1998, pp. 23 – 37.
80. G. Clooth, W. Burchard, *Jahresberichte Steinzerfall – Steinkonservierung* **4,** Verlag Ernst & Sohn, Berlin 1994, pp. 113 – 116.
81. S. Wagener, U. Schneider, H. Höcker, *Jahresberichte Steinzerfall – Steinkonservierung* **3,** Verlag Ernst & Sohn, Berlin 1993, pp. 73 – 83.
82. S. Lotzmann, B. Schwamborn, *Jahresberichte Steinzerfall – Steinkonservierung* **5,** Verlag Ernst & Sohn, Berlin, 1995, pp. 95 – 103.

83. E. Gerhard-Abozari, H. Höcker, K. Littmann, H. R. Sasse, *Jahresberichte Steinzerfall – Steinkonservierung* **5,** Verlag Ernst & Sohn, Berlin 1995, pp. 83 – 90.
84. E. Wendler, S. Lotzmann, B. Schwamborn, *Jahresberichte Steinzerfall – Steinkonservierung* **6,** Fraunhofer IRB Verlag, Stuttgart 1998, pp. 55 – 63.
85. J. D. Neisel, B. Riecken, *Jahresberichte Steinzerfall – Steinkonservierung* **6,** Fraunhofer IRB Verlag, Stuttgart 1998, pp. 83 – 88.
86. U. Boenkendorf, M. Degenkolb, D. Knöfel, *Jahresberichte Steinzerfall – Steinkonservierung* **5,** Verlag Ernst & Sohn, Berlin 1995, pp. 129 – 137.
87. F. Winnefeld, D. Knöfel, *Jahresberichte Steinzerfall – Steinkonservierung* **6,** Fraunhofer IRB Verlag, Stuttgart 1998, pp. 227 – 236.
88. E. Nägele, *Bautenschutz & Bausanierung* **9** (1986) 122 – 129.
89. R. Snethlage: *Leitfaden Steinkonservierung,* Fraunhofer IRB Verlag, Stuttgart 1997, p. 215.
90. J. Ashurst, F. G. Dimes: *Conservation of Building and Decorative Stone,* **vol. 1,** Butterworth/Heinemann, Oxford – London 1990, p. 193, **vol. 2,** p. 254.
91. H. R. Sasse, R. Snethlage: *Saving Our Cultural Heritage. The Conservation of Historic Stone Structures,* Dahlem Workshop Report, John Wiley & Sons, Chichester – New York 1997, pp. 223 – 243.

Chemical Grouting → Silicates

Health Value Added Foods

Ralf Zink, Nestlé-Research Center, Lausanne, Switzerland

Andrea Pfeifer, Nestlé-Research Center, Lausanne, Switzerland

1. **Introduction** 1
2. **Definitions** 1
3. **Prebiotics** 2
4. **Probiotics** 3
5. **Selection Criteria for Probiotic Strains** 4
6. **Physiological Effects, Clinical Applications, and Safety** 5
7. **Outlook** 7
8. **References** 8

1. Introduction

In Western countries the awareness of the nutrition – health relationship among consumers and in the food industry increased sharply during the last 1990s. Several reasons for this increased awareness can be identified:

- Better knowledge on the nutrition – health relationship
- Increasing incidence of nutrition-related chronic diseases, mainly caused by the increasing average age of the population
- Increased purchasing power of consumers
- Increased knowledge in biological, immunological, and nutritional sciences
- Developments in food processing and food technology

New concepts in nutritional sciences have therefore been developed, e.g., functional foods or health value added foods (HVAF) [1]. HVAF refers to food that provides physiological or health benefits beyond the underlying basic nutrition [2] and is an important development in the prevention of chronic nutrition related diseases. It is associated with a shift of interest from health aspects of the total diet towards the health aspects of individual foods, without neglecting the importance of a balanced total diet [3], [4]. In Europe the potential of these foods was addressed in the European Union Demonstration Project (FAIR-CT96-1028: "Demonstration of Nutritional Functionality of Probiotic Foods"). New food ingredients with potential health benefits include dietary fibers and oligosaccharides (prebiotics); polyunsaturated fatty acids, antioxidants, vitamins, minerals, phytochemicals, and specific microorganisms (probiotics). Much scientific and public interest has focused on these probiotics [5–10]. Almost a hundred years have passed since the introduction of these theories of modulation of the intestinal ecosystem [11]. However, only recently has the scientific basis of probiotics been established and sound clinical studies on some bacterial strains published. The physiological and nutritional properties of some strains are now much better understood. Furthermore, it was possible to verify that some strains are probiotic with documented and well-published examples of their ability to maintain and promote the health of their host [12], [13].

2. Definitions

In order to clear the to some extend any existing degree of confusion how and when the different applied terminologies are used, the definition of several keywords should be introduced first. In market surveys, in the public printing-press and media, or even in the scientific literature, the different terms are sometimes used in place of another, without transferring the implied context. This is even more the case when we consider the excessive use of literal translations between the different languages. This can finally result in differing or unprecise definitions used by the regulatory authorities. In order to avoid this, we want to define some of the basic terms used in this article to contribute to a higher clarity of this text.

Health Value Added Foods. There is no universally accepted definition of HVAF foods so far. Nevertheless, a recent definition covers

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.d12_d01

this wide field to a great extent [1]: "a food can be considered as being functional, the moment it can be satisfactorily demonstrated that it affects beneficially one or more target functions in the host body, beyond an adequate nutritional effect, to improve the state of health or well-being, or reduces the risk of a disease". Not all of these claims are allowed in all countries, but it can neverthless be stated that functional foods have to go beyond the basic functions of enriched or fortified foods (e.g., vitamins, minerals). Consequently, a food containing a probiotic (i.e., a living bacteria) is a functional food. Furthermore, it must be possible to consume this foodstuff in quantities comparable and compatible with a normal, well-balanced, and varied diet.

Prebiotics are fermentable, non-viable food components (dietary fibers) that demonstrate a beneficial effect on the intestinal microflora of the host. This benefit to the host is mediated through a selective stimulation of the growth and/or activity of one or a limited number of species of the colonic microflora. As relevant substances, dietary fibers, lactulose, fructo- and galacto-oligosaccharides (FOS and GOS), and inulin (e.g., chicory extract) have been identified. They must not be hydrolyzed or absorbed in the upper gastrointestinal tract (GIT). They are either produced in-situ by bacteria or they are added exogeneously to the food [14]. In general they must alter the microbiota in the colon to a healthier composition and should induce luminal or systematic effects that are beneficial to host health. Bifidobacteria and/or lactobacilli are good target organisms [14], [15].

Probiotics are living microorganisms which, when ingested as a food supplement in sufficient numbers, exert a beneficial impact on the gastrointestinal balance and therefore a positive health impact that goes beyond the conventional nutritional effect [1], [16–20]. The most commonly used probiotics are lactic acid bacteria, such as lactobacilli or bifidobacteria, which are either added to fermented milk products, are responsible for the fermentation of the milk matrix itself, or are given as a supplement in a lyophilized form. Whilst the classical definition states that probiotics are living microorganisms, there is some new and on-going debate whether the concept should include dead microorganisms, or even bacterial fragments. So far, insufficient scientific data regarding their functionality is only available which would justify such an inclusion.

Synbiotics are a combination of a probiotic microorganism and a prebiotic substance. This mixture should benefit the host by improving the survival and implantation of the selected living bacteria in the gastrointestinal tract, by selectively stimulating the growth and/or by activating the metabolism of one or a limited number of health-promoting bacteria [14]. A combination of a bifidobacterial strain with fructose- or galactose-containing oligosaccharides for example, is a common and feasible approach. Only prebiotic bifidogenic factors have been concisely reported so far.

3. Prebiotics

The view of dietary carbohydrates as simply providing us with energy is outdated [21]. Because of their varied chemical and physical form, the rate and extent to which the different types are digested and absorbed in the small intestine varies. Dietary carbohydrates that escape digestion in the upper gastrointestinal tract form the predominant substrates for bacterial growth in the colon. This prebiotic approach offers a support for those probiotics which have insufficient survival rates and/or difficulties in colonization [22–24]. While some peptides, proteins, and certain lipids are potential prebiotics, nondigestible carbohydrates, including those that contain fructose, glucose, xylose and galactose, have received the most attention [21, 25, 26]. For a food ingredient to be classified as prebiotic, it must

1) Be neither hydrolyzed nor absorbed in the upper part of the gastrointestinal tract
2) Be a selective substrate for one or a limited number of beneficial bacteria commensals in the colon, which are stimulated to grow and/or are metabolically activated
3) Consequently, be able to alter the colonic flora in favor of a healthier composition
4) Induce luminal or systemic effects that are beneficial to host health [14].

Some nondigestible carbohydrates have a number of functional effects on the GI tract

which have been used to validate functional and health claims. These include delayed gastric emptying, modulation of GI tract transit times, improved glucose tolerance, reduced fat and cholesterol absorption via binding of bile salts, increased volume and water-carrying capacity of the intestinal contents, and modulation of microbial fermentation with increased short chain fatty acids (SCFA) production, and decreased pH and ammonia production [15], [25], [27]. The combination of these effects could potentially result in an improved host health by reducing intestinal disturbances (constipation and diarrhoea), cardiovascular disease, and intestinal cancer [15].

According to [15], the carbohydrate group that has received the most attention and research are the oligosaccharides. Fructooligosaccharides (FOS) have been extensively studied. A number of foods such as chicory, garlic, onion, artichoke, and asparagus contain relatively high concentrations of FOS [14]. FOS and other oligosaccharides can also be produced enzymatically, which is advantageous for large-scale commercial production [28]. In 1995, 12 food-grade oligosaccharides were in commercial production: lactulose, galactooligosaccharides, fructo-OS, isomalto-OS, malto-OS, palatinose-OS, glucosyl sucrose, cyclodextrins, soybean-OS, lactosucrose, gentio-OS, and xylo-OS. The prebiotic effect of these oligosaccharides is that they may selectively stimulate growth of bifidobacteria. Only glycosyl-sucrose, malto-OS, and cyclodextrins are not reported as being bifidogenic. According to different authors [29], [30] bifidobacteria seem to have a number of health-promoting properties such as improved intestinal environment, inhibition of pathogenic bacteria, immunomodulation, synthesis of vitamin B, improved calcium absorption, lowering of blood ammonia and cholesterol concentrations, and reported tumor formation in animal models.

Effects of feeding prebiotics to human volunteers. A wide number of studies have investigated the effects of FOS on human gut bacteria [15]. In general, feeding of FOS increases bifidobacteria and lactobacilli, increases SCFA concentrations and decreases clostridia, fusobacteria, bacteroides and pH [29], [31–33]. The selective fermentation of FOS by bifidobacteria has also been confirmed in-vivo in healthy human volunteers [14]. Moreover, when bifidobacteria grow on such substrates, they seemingly do so at the expense of bacteroides, clostridia or coliforms, which are maintained at low levels or may even be reduced [34]. The intake of FOS which have elicited a bifidogenic effect in human studies ranged from 4 to 15 g per day [25], [29], [35]. It is critical that the prebiotic effects are tested in properly controlled volunteer trials. Many bacteria are able to metabolize prebiotics in monoculture. However, the key-issue is their behaviour in a competitive ecosystem, such as occurs when feeding to healthy volunteers [15]. Some prebiotics, as inulin and oligofructose, have been reported to improve the bioavailability of some essential minerals such as calcium [36]. Upon inulin ingestion, the apparent Ca-absorption increased significantly. The increase of Ca-absorption did not negatively affect the absorption of other minerals such as magnesium, iron or zinc.

The advantage of prebiotics being carbohydrate materials is that it should be straightforward to incorporate prebiotics into common foodstuff such as cereals, biscuits, cakes, infant formula, weaning food, etc. without too much disturbing either the sensory qualities or texture of the product.

4. Probiotics

The word "probiotic" (Greek: "for life") has had many definitions in the past [16]. It was first used ca. 1965 to describe substances produced by one protozoan that stimulated the growth of another. Around 1970 it was used to describe tissue extracts that stimulated the growth of another tissue [37]. In 1974, Parker used the term to describe animal feed supplements: organisms and substances that have a beneficial effect on the host animal by contributing to its intestinal microbial balance [37]. Fuller (1989) defined the term probiotics in a way we basically still use nowadays as "a live microbial feed supplement which beneficially affects the host by improving its intestinal microbial balance" [37]. This revised definition stresses the importance of living cells as an essential component of an effective probiotic [38].

This definition was broadened to cover treatments with living bacteria that are not administered by mouth and that do not exert their effect through the intestinal microflora: "a mono- or mixed culture of live microorganisms, applied to man or animal as a fermented product or dried cells, that affects beneficially the host by improving the properties of the indigeneous microflora" [39]. This definition therefore includes probiotic products which exert these benefits in the mouth or gastrointestinal tract (taken as food or supplements), in the upper respiratory tract (administered as an aerosol), or in the urogenital tract (local application) [40].

The historical record of the probiotic properties of living microbial food supplements (particularly lactic acid bacteria) is long. As early as 76 B.C., PLINIO advocated the use of fermented milks in the treatment of various forms of gastrointestinal infections [41]. The effect of probiotics has been investigated since the early 1900s, before the term was even coined. One of the first to advocate that sour milk containing bacteria may have a positive effect on intestinal health and to provide an initial scientific explanation was MECHTNIKOFF in 1907 [11]. He proposed that the consumption of large quantities of yoghurt containing *Lactobacillus* spp. would result in the replacement of toxin-producing bacteria normally present in the intestine and thus result in better health and an increased lifespan. TISSIER (1906) had already proposed the administration of bifidobacteria to infants suffering from diarrhoea, believing that these bacteria would replace "putrefactive bacteria responsible for gastric upset, while re-establishing themselves as the dominant intestinal microorganism" [33], [42]. METCHNIKOFF believed that careful management of the gut microflora could exert beneficial effects on overall human health. In order to "manage" the gut microflora, he was particularly attracted by "exotic foods like kefir, yoghurt and sour milk" [43]. Nevertheless, it is impossible to state which bacterial species were present in these products at that time as they were still prepared in a very folkloristic way. Consequently, the majority of early research was (and is still) performed with microorganisms that are used to produce fermented milk products such as yoghurt. However, even in the early 1920s, researchers suggested that intestinal isolates might have better, more distinctive effects than non-intestinal organisms [16]. This idea is unfortunately not even today always considered in the selection of the wide variety of existing probiotic strains [15]. Recent well-documented clinical studies support health claims for specific bacterial strains, including *Lactobacillus johnsonii* La1 [44], *Lb. rhamnosus* GG [45], *Lb. casei* Shirota strain, *Lb. acidophilus* NCFB 1487, and *Bifidobacterium lactis* Bb12.

5. Selection Criteria for Probiotic Strains

There are a number of important properties that a bacterium must possess to make it a desirable candidate to become a probiotic [12], [40], [46], [47]. The bacterium must be able to survive the acidic conditions of the upper GIT, and proliferate in and/or colonize the intestine [48]. The absence of pathogenic, mutagenic, toxic, or carcinogenic effects on the host of its fermentation products or cell compounds must be assured. In addition, the probiotic bacteria could be antagonistic towards microorganisms involved in carcinogenicity and pathogenicity [49]. Furthermore, its genetic stability and the absence of a plasmid-transfer mechanism must be assured by means of conjugation experiments. Probiotic strains for humans should be of human origin, because some health-promoting effects might be species and origin dependant. They should be normal inhabitants of the intestinal tract, since this eases their adaption to the digestive system.

A classification of probiotic microorganisms and their infection potential follows [8]:

Lactobacillus	mainly nonpathogens, some opportunistic infections possible (usually in severely immunocompromised patients)
Streptococcus	oral streptococci mainly non-pathogenic (including *Str. thermophilus*); some may cause opportunistic infections
Enterococcus	some strains are opportunistic pathogens with hemolytic activity and even transferable antibiotic resistance (e.g., vancomycin)
Bifidobacterium	mainly non-pathogens, some isolated cases of human infections
Saccharomyces	mainly non-pathogens, some isolated cases of human infections

To be industrially applicable, the probiotic strain must be suitable for technological processes (easy to reproduce) and it must remain

viable during processing and storage in the final product or as a supplement.

Characteristics expected of potential probiotic strains are:

- Microbiological and safety parameters
  - Accurate taxonomic identification
  - Normal inhabitants of the target species (human origin)
  - Non-toxic and non-pathogenic and non-allergic
  - Genetically stable
  - Capable of survival, proliferation, and metabolically active at the target site
  - Potentially resistant to bacteriocins, acids, and other antimicrobials produced by the residing microflora
  - Resistant to bile salts
  - Resistant to acids, e.g., gastic juice
- Health benefits
  - Adherence and colonization potential preferred
  - High populations ($10^6 - 10^8$) per uptake
  - Antagonistic towards pathogenic/cariogenic bacteria
  - Production of antimicrobial substances (e.g. bacteriocins, hydrogen peroxide, organic acids)
  - Able to compete with normal microflora in a transient and/or permanent manner
  - Immunomodulatory
  - Able to exert clinically documented health benefits
- Industrial parameters
  - Stability of desired characteristics during culture preparation, storage, and delivery
  - Amenable to production processing: adequate growth, recovery, concentration, freezing, dehydration, lyophilization, storage, and distribution
  - Provision of desirable organoleptic qualities when included in the fermentation process

6. Physiological Effects, Clinical Applications, and Safety

A wide variety of beneficial functions of probiotic strains [59–61] have been reported or were theorized [62] (Table 1).

Table 1. Probiotic bacterial and yeast strains and their reported effects [12], [50–58]

Reported effects	Species [a]
Stimulation/modulation of the immune system [b]	*Lb. acidophilus, Lb. johnsonii, Lb. casei, Lb. plantarum, Lb. rhamnosus*
Balancing the intestinal microflora	*Lb. acidophilus, Lb. johnsonii, Lb. casei, Bif. lactis, Bif. bifidum*
Reduced fecal enzymes [c]	*Lb. acidophilus, Lb. casei, Lb. gasseri*
Antitumor	*Lb. acidophilus, Lb. casei, Lb. gasseri, Lb. plantarum, Bif. infantis, Bif. adolescentis, Bif. longum, Bif. lactis, Bif. bifidum*
Prevention of travellers diarrhoea	*Sacch. boulardii*
Prevention of rotavirus diarrhoea	*Lb. rhamnosus, Bif. bifidum, Bif. lactis*
Prevention of *Clostridium difficile* diarrhoea	*Lb. rhamnosus, Sacch. boulardii*
Prevention of other diarrhoea [d]	*Lb. acidophilus, Lb. rhamnosus, Lb. reuteri, Bif. bifidum, Bif. lactis*
Anti-Helicobacter activity	*Lb. johnsonii, Lb. acidophilus*

[a] *Lb = Lactobacillus, Bif. = Bifidobacterium, Sacch. = Saccharomyces;* only specific strains exhibit these characteristics; dairy strains of non-human origin and questionable high dosages (e.g., 500 g yoghurt per day) were not considered, due to unrealistic and unconventional dosages. [b] A variety of probiotic effects on the immune system have been reported. [c] Primarily those that convert non-toxic compounds into mutagens or carcinogens. [d] Including diarrhoea of viral origin and that caused by chemotherapy and antibiotic treatment.

Knowledge of the biological and clinical effects of various strains of probiotics is summarized in the following:

- Well established
  - Modulation of the immune system
  - Reduction of some diarrhoeal disorders (rotavirus, antibiotic effect, *E. coli* enterotoxin)
  - Anti *Helicobacter pylori* activity
  - Alleviation of lactose intolerance
  - Prevention of urogenitial infections
  - Alleviation of constipation
- Reasonably established
  - Improved resistance to intestinal infections and disorders
  - Change in fecal enzymatic activities
  - Improved nutritional value of food
  - Maintenance of normal intestinal and urogenital microflora
  - Protection against traveller's and infantile diarrhoea
  - Reduction of antibiotic-induced diarrhoea

- Suggested evidence: not yet clearly demonstrated and remains doubtful so far (only in-vitro or animal model data)
  - Reduction of serum cholesterol levels
  - Antitumour and anticancer activity
  - Vitamin production
  - Influence on hormone regulation
  - Prevention of hypercholesterolemia
  - Protection from colon/bladder cancer
  - Reduction of side effects of hepatic encephalopathy
  - Aid in hypo- and hyperchlorohydria
  - Prevention of osteoporosis

Adherence to human intestinal epithelial cells is considered a very important characteristic, since any probiotic strain which can become established, even only temporarily, is likely to have a competitive advantage and a greater or more prolonged probiotic effect. Adhesion to enterocytes and/or fecal colonization are important selection criteria for immunomodulation [63–66].

An important property of probiotic strains is their antagonistic activity against pathogenic strains by competitive exclusion, bacterial aggregation, or production of antimicrobial substances, including organic acids, hydrogen peroxide, and bacteriocins [67–70]. Although there is evidence that probiotic LAB reduce the risk of certain cancers in animal models, there is insufficient clinical data so far available to extrapolate these results to humans [71].

Interestingly, it was also discovered that the *Lb. johnsonii* La1 supernatant could significantly inhibit the *in-vitro* and *in-vivo* growth of the gastric pathogen *Helicobacter pylori* and other pathogens [53], [57]. A double-blind, controlled clinical study with *H. pylori* positive patients confirmed this result in the *in-vivo* situation, raising the question if and how far functional metabolites or bioactive molecules are involved [52], [53].

A critical evaluation of probiotics in the treatment of acute diarrhoea is difficult to prove in clinical practice, since many different probiotic microorganisms are used in different forms and in different combinations (yoghurt, milk, dried, etc.) [72]. The yeast *S. boulardii* is on of the probiotics with the most convincing and reproducible double-blind studies for an antidiarrhoea treatment so far. However, in some studies on bacteria it is not completely clear if the experiments were performed with viable cells or with their lysates. The effectiveness of bacteria in this aspect has to be more effectively demonstrated in well-designed randomized, placebo-controlled and double-blind studies before their large-scale use can be recommended [73], [74].

Requirements for good clinical studies with probiotic bacteria are [12]:

- Defined and well-characterized bacterial strains
- Well-defined studies and well-designed clinical trials
- Double-blind; placebo-controlled; randomized
- Results confirmed by different groups
- Publication in peer-reviewed journals

Probiotics are viable microorganisms, used in foods, especially fermented dairy products, but also in pharmaceutical preparations. The development of new probiotic strains aims at more active beneficial organisms [75], [76]. In the case of these novel microorganisms the question of safety and their risk-to-benefit ratio have to be assessed. Lactic acid bacteria (LAB) and Bifidobacteria have a long history of safe use in foods. Members of the genera *Lactobacillus* and *Bifidobacterium* are most commonly given the generally recognized as safe status (GRAS), whilst the genera *Streptococcus* and *Enterococcus*, and some other genera of LAB contain some opportunistic pathogens. Lactic acid bacteria are intrinsically resistant to many antibiotics, but in many cases these resistances are not transmissible, and these species are also sensitive to many clinically used antibiotics even in the case of LAB-associated opportunistic infections. Therefore, no particular safety concern is associated with intrinsic types of resistance. Plasmid-associated antibiotic resistance, however, which occasionally occurs, is another matter because of the possibility of the resistance spreading to other, more harmful species and genera. The transmissible enterococcal resistance against glycopeptide antibiotics (e.g., vancomycin) is particularly noteworthy, as vancomycin is one of the last effective antibiotics in the treatment of certain multidrug-resistant pathogens (e.g., staphylococci). New species and more specific strains of probiotic bacteria were identified in the last five years. Prior to incorporating these probiotic strains into

products their efficacy should be carefully addressed, and a case-by-case evaluation as to whether they share the safety status of traditional food-grade organisms should be made [8], [75], [77]. A theoretical side-effect of all probiotics is the risk of translocation into the bloodstream. All of the probiotics which are widely used in Western Europe and the United States appear to be safe. Only about 15 cases (mostly *S. boulardii* and *L. rhamnosus*) in a period of 10 years, during which millions of consumptions of probiotic-containing products or prescribed treatment of diarrhoea with yeasts, have been published in well-documented clinical cases in peer-reviewed scientific and clinical journals.

7. Outlook

Pre-, pro- and synbiotics are able to modify the intestinal microbial composition in the gastrointestinal tract. Dietary carbohydrates can no longer be viewed as simple energy providers only, the diversity of physiological and health effects need to be considered. It is only on this basics possible to understand the importance for health of probiotics for the human consumer [78]. Among the natural non-digestible oligosaccharides that fulfill the criteria of a prebiotic, FOS and inulin are the only products presently generally recognized and used as food ingredients. Other compounds with experimental evidence are, e.g., transgalactosylated disaccharides [79–82] and soybean oligosaccharides [83], [84].

The scientific literature reveals that probiotics offer a large number of theoretical advantages in the promotion of health beyond providing only basic nutritional aspects, justifying their attribute of being a HVAF. They offer a possibly partial solution to the problem of the increasing resistance of bacteria to antibiotics, and have the advantage of possessing multiple ways to express their effect. This makes it very unlikely that resistance to probiotics and their implied benefit might ever develop. They do not change the already existing intrinsic resistance of the intestinal colonizing microflora.

Numerous clinical trials have demonstrated that specific probiotic bacterial strains are able to deliver their health benefit, such as balancing the intestinal microflora, modulating the immune system or preventing gastrointestinal tract infections [85–88]. Many validated probiotic lactobacilli- or bifidobacteria-strains found naturally in the human intestinal tract have an excellent potential for a further inclusion into the functional food concept [44], [89]. In order to support health claims for pre- and probiotics, more collaborative data has to be created to associate changes in gut bacterial populations with the specific physiological aspects and changes in humans. Additionally, a better understanding of how pre- and probiotics cause changes in the microbial community is urgently required [31].

Particularly desirable effects described for various lactic acid bacteria are their activity against pathogens and strengthening the gut mucosal barrier [45], immune modulation [65], [66], anti-allergenic effects [90], anti-mutagenic and anti-tumorigenic activity, which have to be shown in more detail [50], [91], and further nutritional advantages [78]. In particular, the underlying molecular mechanisms for these probiotic properties need further scientific evaluation. Only the thorough demonstration of clinical and molecular effects by selected bacterial strains will give the probiotic concept its inherent credibility among the scientific community and consumers [13]. Given the ease of use, the consumer will be more attracted towards the concepts the more definitive health aspects are demonstrated, reported and proven [15]. The approach offers much potential in the prophylactic management of gastrointestinal diseases (such as inhibition of pathogens, prevention of bowel cancer or irritable bowel syndromes) and more systemic effects (e.g., potential reduction of blood lipids, potential hormonal regulation).

Future research on probiotic bacteria will most likely still center on the selection of new and more-specific strains of lactic acid bacteria (LAB) and other bacteria to promote the well-being of the host and which will fit into the definition of HVAF. Different regions of the gastrointestinal tract require different probiotic bacteria. Knowledge of the intestinal microflora is necessary in order to understand the impact of a probiotic strain on the microbial community in the intestinal tract [13], [31]. Dynamic models of the stomach, small intestine or colon to simulate the in-vivo gastrointestinal environment have been developed and validated and will provide a link between conventional in-vitro meth-

ods and in-vivo clinical and human studies. Furthermore, it will be necessary to select disease-specific strains, which could, for example, act against rotavirus diarrhoea or gastritis caused by *Helicobacter pylori*. Applications of molecular methods to the study of intestinal microbial ecology will help to achieve these research goals. Oligonucleotide probes targeting 16S ribosomal RNA have been used to study the microbial ecology [92], [93]. Other molecular methods used in molecular microbiology to determine unequivocally the effects of pre-, pro- and synbiotics on gut microbial ecology and to monitor microbial population changes include DNA-fingerprinting, 16S-rRNA restriction/amplified-fragment length polymorphism (RFLP/AFLP), and pulsed-field-gel-electrophoresis (PFGE).

Today, most probiotic strains are used in dairy products. Present studies indicate that new probiotic functional foods will include infant formulas, baby foods, fermented fruit juices, fermented soy products, fermented oatmeal without milk, and also disease-specific clinical foods containing viable and physiologically active probiotic bacteria. All these products will have to possess good stability and in many cases prolonged shelf-life. For such development to occur, it is important that the requirements for good clinical studies are followed even more carefully. In addition, the effects of mixtures of different probiotic strains should be investigated. With carefully controlled studies on clearly defined targets and well-selected strains the future will provide probiotic bacteria even more tailor-made for specific diseases and their subsequent prevention [12]. Information concerning both the individual bacteria and the complex microbial populations which are of intestinal origin will add to our overall knowledge of the ecology and activity of microorganisms in the gut intestinal tract and their relationship to our well-being [13].

8. References

1. A. T. Diplock et al.: Scientific concepts of functional foods in Europe, consensus document. *FF-27-de98, Bruxelles: ILSI Europe*, pp. 17 – 79.
2. Y. Delneste, A. Donnet-Hughes, E. Schiffrin: "Functional Foods: Mechanisms of action on immunocompetent cells," *Nutr. Rev.* **56** (1998) no. 1, S93 – S98.
3. M. E. Sanders.: "Summary of the conclusions from a consensus panel of experts on health attributes of lactic cultures: Significance of fluid milkproducts containing cultures," *J. Dairy Sci.* **76** (1993) 1819 – 1828.
4. M. E. Sanders.: "Overview of functional foods: emphasis on probiotic bacteria," *Int. Dairy J.* **8** (1998) 341 – 347.
5. P. Kaufmann: "Functional Food: The monitoring of probiotic strains by modern DNA based technology," *Bio-World; Lebensmitteltechnol.* **4** (1997) 12 – 15.
6. J. K. Collins, G. Thornton, G. O. Sullivan: "Selection of probiotic strains for human applications," *Int. Dairy J.* **8** (1998) 487 – 490.
7. P. Conway: "Lactobacilli: fact and fiction," in R. Grubb, T. Midtvedt, E. Norin (eds.): *The regulatory and protective role of the normal microflora,* Stockton Press, New York 1989, pp. 263 – 281.
8. S. E. Salminen, A. C. Ouwehand, E. Isolauri: "Clinical applications of probiotic bacteria," *Int. Dairy J.* **8** (1998) 563 – 572.
9. G. W. Tannock: "Probiotic properties of lactic acid bacteria: Plenty of scope for fundamental R&D," *Trends Biotechnol.* **15** (1997) 270 – 274.
10. G. W. Tannock: *Probiotics: A critical review,* Horizon scientific press, Norfolk, U.K. 1999
11. E. Mechtnikoff: *The prolongation of life,* Heinemann, London 1907.
12. Y.-K. Lee, S. Salminen: "The coming age of probiotics," *Trends Food Sci. Technol.* **6** (1995) 241 – 244.
13. E. E. Vaughan, B. Mollet: "Probiotics in the new millenium," *Nahrung* **43** (1999) 148 – 153.
14. G. R. Gibson, M. B. Roberfroid: "Dietary modulations of the human colonic microbiota: Introducing the concept of prebiotics," *J. Nutr.* **125** (1995) 1401 – 1412.
15. C. J. Ziemer, G. R. Gibson: "An overview of probiotics, prebiotics and synbiotics in the functional food concept. Perspectives and future strategies," *Int. Dairy J.* **8** (1998) 473 – 479.
16. R. Fuller: "History and development of probiotics", in R. Fuller (ed.): *Probiotics: The scientific basis,* Chapman & Hall, New York 1992, pp. 1 – 8.
17. J. A. Vanderhoof, R. J. Young: "Use of probiotics in childhood gastrointestinal

disorders," *J. Pediatr. Gastroenterol. Nutr.* **27** (1998) 323 – 332.
18. J. A. Vanderhoof, R. J. Young, N. Murray, S. S. Kaufman: "Treatment strategies for small bowel bacterial overgrowth in short bowel syndrome," *J. Pediatr. Gastroenterol. Nutr.* **27** (1998) 155 – 160.
19. S. E. Gilliland: "Health and nutritional benefits from lactic acid bacteria," *FEMS Microbiol. Rev.* **7** (1990) 175 – 188.
20. R. Havenaar, J. H. J. Huis in't Veld: "Probiotics: a general view," in B. J. B. Wood (ed.): *The lactic acid bacteria, Volume 1: The lactic acid bacteria in health and disease,* Chapman & Hall, New York 1992, pp. 209 – 224.
21. N. Delzenne, M. Roberfroid: "Physiological effect of non-digestible oligosaccharides," *Lebensm. Wiss. u. Technol.* **27** (1994) 1 – 6.
22. P. Kontula, J. Jaskari, L. Nolelt, I. De Smet, A. von Wright, K. Poutanen, T. Mattila-Sanholm: "The colonization of a simulator of the human intestinal microbial ecosystem by a oprobiotic strain fed on a fermented oat bran product: effects on the gastrointestinal microflora," *Appl. Microbiol. Biotechnol.* **50** (1998a) 246 – 252.
23. P. Kontula, A. von Wright, T. Mattila-Sanholm: "Oat bran beta-gluco- and xylo-oligosaccharides as fermentative substrates for lactic acid bacteria," *Int. J. Food Microbiol.* **45** (1998b) 163 – 169.
24. M. L. Johansson, G. Molin, B. Jeppson, S. Nobaek, S. Ahrne, S. Bengmark: "Administration of different Lactobacillus strains in fermented oatmeal soup. In-vivo colonization of human intestinal mucosa and effect on the indigenous flora," *Appl. Environ. Microbiol.* **59** (1993) 15 – 20.
25. M. B. Roberfroid: "Functional effects of food components and the gastrointestinal system: chicory fructooligosaccharides," *Nutr. Rev.* **54** (1996) 538 – 542.
26. M. B. Roberfroid, J. A. E. van Loo, G. R. Gibson: "The bifidogenic nature of chicory inulin and its hydrolysis products," *J. Nutr.* **128** (1998) 11 – 19.
27. J. H. Cummings, "Short chain fatty acids (SCFA)," in G. R. Gibson, G. T. MacFarlane (eds.): *Human colonic bacteria: Role in Nutrition, Physiology and Pathology,* CRC-Press, Boca Raton, FL 1995, pp. 101 – 130.
28. R. G. Crittenden, M. J. Payne: "Production, properties and applications of food-grade oligosaccharides," *Trends Food Sci. Techno.* **7** (1996) 353 – 361. Crowell, D.C. (1998). Microbial analysis of the human intestinal flora after feeding *Lactobacillus acidophilus*. M.Sc.-thesis, North Carolina State University.
29. G. R. Gibson, A. Willems, S. Reading, M. D. Collins: "Fermentation of non-digestible oligosaccharides by human colonic bacteria," *Proc. Nutr. Soc.* **55** (1996) 899 – 912.
30. T. Yaeshima: "Benefits of bifidobacteria to human health," *Bull. IDF 313: Oligosaccharides and probiotic bacteria* (1996) 36 – 42.
31. R. Fuller, G. R. Gibson: "Modification of the intestinal microflora using probiotics and prebiotics," *Gastroenterology* **32** (1997) Suppl. 222, 28 – 31.
32. G. R. Gibson, E. B. Beatty, X. Wang, J. H. Cummings: "Selective stimulation of bifidobacteria in the human colon by oligofructose and inulin," *Gastroenterology* **108** (1995) 975 – 982.
33. M. G. O'Sullivan: "Metabolism of bifidogenic factors by gut flora – an overview," *Bull. IDF 313: Oligosaccharides and probiotic bacteria* 1996, pp. 23 – 30.
34. X. Wang, G. R. Gibson: "Effects of the in-vitro fermentation of oligofructose and inulin by bacteria growing in the human large intestine," *J. Appl. Bacteriol.* **75** (1993) 373 – 380.
35. R. K. Buddington, C. H. Williams, S. C. Chen, S. A. Witherly: "Dietary supplementation of neosugars alters the fecal flora and decreases activities of some reductive enzymes in human subjects," *Americ. J. Clin. Nutr.* **63** (1996) 709 – 716.
36. A. Franck: "Prebiotics stimulate calcium absorption: a review," *Milchwiss.* **53** (1998) no. 8, 427 – 429.
37. R. Fuller: "Probiotics in man and animals," *J. Appl. Bacteriology.* **66** (1989) 365 – 367.
38. M. G. O'Sullivan, G. Thornton, G. G. O'Sullivan, J. K. Collins: "Probiotic bacteria: Myth or reality?" *Trends Food Sci. Technol.* **3** (1992) 309 – 314.
39. J. H. J. Huis in't Veld, R. Havenaar, P. Marteau: "Establishing a scientific basis for probiotic R&D," *Tibtech.* **12** (1994) 6 – 8.
40. R. Havenaar, B. ten Brink, J. H. J. Huis in't Veld: "Selection of strains for probiotic use," in R. Fuller (ed.): *Probiotics: The Scientific Basis,* Chapman & Hall, New York 1992, pp. 209 – 224.

41. V. Botazzi: "Food and feed production with microorganisms," in G. Reed (ed.): *Biotechnology* **vol. 5**VCH, Weinheim 1983, 315 – 363.
42. H. Tissier: "Bifidobacteria and the benefit," *C. R. Soc. Biol.* **60** (1906) 359 – 361.
43. "Gut microflora and health," Nestlé's Food and Nutrition Backgrounder, Nestlé Nutrition Communication, stratetic business division, Vevey 1999.
44. R. Reniero: "Lactobacillus johnsonii La1. Probiotic research and design of new food," *Gastroenterologia, Suppl.* **11** (1998) 45 – 47.
45. S. E. Salminen, E. Isolauri, E. Salminen: "Clinical use of probiotics for stabilizing the gut mucosal barrier: Successful strains and future challenges," *Antonie van Leeuwenhoek* **70** (1996) 347 – 358.
46. A. Donnet-Hughes, F. Rochat, P. Serrant, J. A. Aeschlimann, E. J. Schiffrin: "Modulation of non-specific mechanisms of defense by lactic acid bacteria: effective dose," *J. Dairy Sci.* **82** (1998) 863 – 869.
47. T. R. Klaenhammer: "Microbiological considerations in selection and preparation of Lactobacillus strains for use as dietary adjuncts," *J. Dairy Sci.* **65** (1982) 1339 – 1349.
48. D. Granato, F. Perotti, I. Masserey, M. Rouvet: "Cell surface-associated lioteichoic acid acts as an adhesion factor for attachment of Lactobacillus johnsonii La1 to human enterocyte-like Caco-2 cells," *Appl. Environ. Microbiol.* **65** (1999) 1071 – 1077.
49. M. C. Johnson, B. Ray, T. Bhowmink: "Selection of Lactobacillus acidophilus strains for use in "acidophilus products"," *Antonie van Leeuwenhoek* **53** (1987) 215 – 231.
50. A. Adachi in B. J. B. Wood (ed.): *Lactic acid bacteria,* **vol. I,** Elsevier Sci. Publ., London 1992.
51. G. W. Tannock: "Role of probiotics," in G. R. Gibson, G. T. MacFarlane (eds.): *Human colonic bacteria: Role in Nutrition, Physiology and Pathology,* CRC Press, Boca Raton, FL 1995, pp. 257 – 271.
52. M. H. Coconnier, V. Lievin, E. Hemery, A. L. Servin: "Antagonistic activity against Helicobacter infection in-vitro and in-vivo by the human Lactobacillus acidophilus strain LB," *Appl. Environ. Micro.* **64** (1998) 4573 – 4580.
53. P. Michetti, G. Dorta, P. H. Wiesel, D. Brassart, E. Verdu, M. Herranz, C. Felley, N. Porta, M. Rouvet, A. L. Blum, I. Corthesy-Theulaz: "Effect of whey-based culture supernatant of Lactobacillus acidophilus (johnsonii) La1 on Helicobacter pylori infection in humans," *Digestion* **60** (1999) 203 – 209.
54. M. Aleander, R. Satokari, R. Korpela, M. Saxelin, T. Vilpponen-Salmela, T. Mattila-Sandholm, A. von Wright: "Persistance of colonization of human colonic mucosa by a probiotic strain, Lactobacillus rhamnosus GG, after oral consumption," *Appl. Environ. Microbiol.* **65** (1999) no. 1, 351 – 354.
55. M. R. Aleander, R. Korpela, M. Saxelin, T. Vilpponen-Salmela, T. Mattila-Sandholm, A. von Wright: "Recovery of Lactobacillus rhamnosus GG from human colonic biopsis," *Lett. Appl. Micribiol.* **24** (1997) 361 – 364.
56. R. Targan, F. Shanahan *Inflammatory bowel disease from bench to bedside,* Williams & Wilkins, Baltimore, USA 1994.
57. E. J. Schiffrin, D. Brassart, A. L. Servin, F. Rochat, A. Donnet-Hughes: "Immune modulation of blood leucocytes in man by lactic acid bacteria: criteria for strain selection," *American J. Clin. Nutr.* **66** (1997) 5155 – 5205.
58. S. Hudault, V. Lievin, M.-F. Bernet-Cambard, A.-L. Servin: "Antagonistic activity in-vitro and in-vivo by Lactobacillus casei (strain GG) against Salmonella typhimurium C5 infection," *Appl. Environ. Microbiol.* **63** (1997) 513 – 518.
59. H. L. Dupont: "Prevention of diarrhoea by the probiotic, Lactobacillus GG," *J. Pediatr.* **134** (1999) 1 – 2.
60. H. S. Gill: "Stimulation of the immune system by lactic cultures," *Int. Dairy J.* **8** (1998) 535 – 544.
61. B. R. Goldin: "Health benefits of probiotics," *British J. Nutr. Suppl.* **2** (1998) S203 – S207.
62. P. Marteau, J. C. Rambaud: "Potential of using lactic acid bacteria for therapy and immunomodulation in man," *FEMS Micro. Rev.* **12** (1993) 207 – 220.
63. M. Miettinen et al.: "Lactobacilli and streptococci induce interleukin-12 (IL-12), IL-18, and gamma-interferon production in human peripheral blood mononuclear cells," *Infect. Immun.* **66** (1998) no. 12, 6058 – 6062.
64. A. Pfeifer, J.-P. Rosat: "Probiotics in alimentation: clinical evidence for their enhancement of the natural immunity of the gut," in P. Hanson, R. Yolken (eds.): *Probiotics, Other Nutritional Factors, and*

Intestinal Microflora, **vol. 42, Nestlé Nutrition workshop series,** Raven Publ., PA 1999.
65. E. J. Schiffrin, D. Haller, Y. Delneste, A. Donnet-Hughes, S. Blum: "Probiotics and intestinal immune functions," *Gastroenterology International* **11** (1998) 116 – 121.
66. E. J. Schiffrin, F. Rochat, H. Link-Amster, A. Donnet-Hughes: "Immunomodulation of human blood cells following infection of lactic acid bacteria," *J. Dairy Sci.* **78** (1998) 491 – 497.
67. M. F. Bernet-Camard et al.: "The human Lactobacillus acidophilus strain La1 secretes a non-bacteriocin antibacterial substance(s) active in vitro and in vivo," *Appl. Environ. Micro.* **63** (1997) 2747 – 2753.
68. T. R. Klaenhammer: "Genetics of intestinal lactobacilli," *Int. Dairy J.* **5** (1995) 1019 – 1058.
69. S. E. Lindgren, W. J. Dobrogosz: *FEMS Microbiol. Rev.* **87** (1990) 149 – 164.
70. R. J. Spencer, A. Chesson: *J. Appl. Bacteriol.* **77** (1994) 215 – 220.
71. L. Pelto, E. Isolauri, E.-M. Lilius, J. Nuutila, S. Salminen: "Probiotic bacteria down-regulate the milk-induced inflammatory response in milk-hypersensitive subjects but have an immunostimulatory effect in healthy subjects," *Clin. Exper. Allergy* **28** (1998) 1474 – 1479.
72. Huis in't Veld, P. Marteau: "The role of LAB in relation to human health: Progress over the last three years," *Actes du Colloques LACTIC '97*. Caen, Sept. 10 – 12 1997, pp. 13 – 30.
73. E. Isolauri, M. Kaila: "Lactic acid bacteria to cure or prevent infant diarrhoea," *Actes du Colloques LACTIC '97*. Caen, Sept. 10 – 12 1997, pp. 131 – 137.
74. C. Roffe: "Biotherapy for antibiotic-associated and other diarrhoea," *J. Infection.* **32** (1996) 1 – 10.
75. S. Salminen, A. von Wright: "Current probiotics – safety assured?" *Microbiol. Ecol. Health Dis.* **10** (1998) 68 – 77.
76. S. Salminen et al.: "Demonstration of safety of probiotics – a review," *Int. J. Food Microbiol.* **44** (1998) 93 – 106.
77. B. W. Wolf, D. G. Garleb, I. Casas: "Safety and tolerance of Lactobacillus reuteri in healthy adult male subjects," *Microbiol. Ecology. Health Dis.* **8** (1995) 41 – 50.
78. J. H. Cummings, G. T. Mc Farlane: " A new look at dietary carbohydrates: chemistry, physiology and health," *Europ. J Clin. Nutr.* **51** (1997) 417 – 423.
79. M. Ito et al.: "Effect of administration of galacto-oligosaccharides on the human faecal microflora, stool, weight and adbominal sensation," *Microb. Ecol. Health Dis.* **3** (1990) 285 – 292.
80. M. Ito et al.: "Effects of transgalactosylated disaccharides on the human intestinal microflora and their metabolism," *J. Nutr. Sci. Vitaminol.* **39** (1993) 279 – 288.
81. I. R. Rowland: "Metabolic interactions in the gut," in R. Fuller (ed.): *Probiotics. The Scientific Basis,* Chapman & Hall, London 1992, pp. 29 – 53.
82. R. Tanaka et al.: "Effects of administration of TOS and Bifidobacterium breve 4006 on the human fecal flora," *Bifidobacteria Microflora* **2** (1983) 17 – 24.
83. K. Hayakawa et al.: "Effects of soybean oligosaccharides on human faecal microflora," *Microb. Ecol. Health Dis.* **3** (1990) 293 – 303.
84. Y. Saito, T. Takano, I. Rowland: "Effects of soybean oligosaccharides on the human gut microflora in in-vitro culture," *Microb. Ecol. Health Dis.* **5** (1992) 105 – 110.
85. A. O. Ouwehand, S. Salminen: "The health benefits of cultured milk products with viable and non-viable bacteria," *Int. J. Dairy* **8** (1998) 749 – 758.
86. S. Salminen et al.: "Functional food science and gastrointestinal physiology and function," *Brit. J. Nutr. Suppl.* **1** (1998) S147 – S171.
87. A.-V. Shornikova, I. A. Casas, E. Isolauri, H. Mykkänen, T. Vesikari: "Lactobacillus reuteri as a therapeutic agent in acute diarrhoea in young children," *J. Ped. Gastroent. Nutr.* **24** (1997) 399 – 404.
88. A.-V. Shornikova, I. A. Casas, H. Mykkänen, E. Salo, T. Vesikari: "Bacteriotherapy with Lactobacillus reuteri in rotavirus gastroenterology," *Pediatr. Infect. Dis. J.* **16** (1997) 1103 – 1107.
89. M. Saxelin, T. Pessi, S. E. Salminen: "Fecal recovery following oral administration of Lactobacillus strain GG (ATCC 53103) in gelatine capsules to healthy volunteers," *Int. J. Food Micro.* **25** (1995) 199 – 203.
90. H. Majamaa, E. Isolauri: "Evaluation of the gut mucosal barrier: evidence for increased antigen transfer in children with atopic eczema," *J. Allergy Clin. Immunol.* **97** (1996) 985 – 990.
91. A. Hosono, H. Kitazawa, T. Yamaguchi: "Probiotic aspects," in R. Fuller (ed.):

Probiotics 2: Applications and practical aspects, Chapman & Hall, Cambridge 1997, pp. 89 – 132.
92. J. Dore, R. I. Mackie: "Microbial diversity in gut ecosystems," in *Evolution of the Rumen Microbial Ecosystem Symposium*, Aberdeen 1997, pp. 28 – 33.
93. C. J. Ziemer et al.: "Evaluation of ruminal microbial population changes during acidosis using ribosomal RNA-targeted DNA-probes," *J. Dairy Sci.* **79** (1996) Suppl. 1, 141.

Heart Drugs → **Antiarrhythmic Agents**
Heart Drugs → **Cardiovascular Drugs**
Heart Drugs → **Coronary Therapeutics**
Heart Drugs → **β-Receptor Blocking Agents**

Membrane Reactors

José G. Sanchez Marcano, Institut Européen des Membranes, UMII, cc 47, Montpellier, France

Theodore T. Tsotsis, Department of Chemical Engineering, University of Southern California, Los Angeles, CA 90089, United States

1. Introduction 1
2. Catalytic Membrane Reactors 3
2.1. Dehydrogenation Reactions 3
2.2. Hydrogenation Reactions 3
2.3. Oxidation Reactions 4
2.4. Other High-Temperature Applications 6
3. Pervaporation Membrane Reactors . 9
4. Membrane Bioreactors 9
4.1. Whole-Cell MBR Applications 9
4.2. Enzymatic MBRs 11
4.3. Environmental Applications 13
5. Conclusions 14
6. References 15

1. Introduction

Membrane reactors (MRs) are finding use today in the petrochemical, water-treatment, food, and pharmaceutical industries. These reactor systems belong to a class of processes termed reactive separations, which combine reaction and separation into a single operation. The prime driver for such applications is the promise for savings in processing costs. In MRs the separation task is carried out by a membrane. The earlier applications of MR were in the catalysis field. Here combining the reaction and separation steps, in addition to providing process simplicity, also results in enhanced reactor selectivity and/or yield. Examples of such applications are equilibrium-limited reactions (e.g., hydrocarbon dehydrogenations), for which the continuous removal of a reaction product (e.g., hydrogen) often increases the reactor yield. Other examples include partial oxidation and esterification reactions. Some of the most promising MR applications today are in biotechnology, where MRs are used in industrially important biological reactions. These include fermentations used for the production of amino acids, antibiotics, and other fine chemicals. During such processes the membrane eliminates reaction metabolites that inhibit biological function and thus help to maintain high reactor productivity. Membranes are also increasingly being utilized in MRs as hosts for the immobilization of bacteria, enzymes, or animal cells for the production of highly value added chemicals. MRs are also finding increasing application in the treatment of contaminated air and water streams.

MR processes either integrate two physically distinct units – the reactor and the membrane module (Figure 1 a) – or incorporate both functions into a single unit (Figure 1 b). The compact design of Figure 1 b promises savings in capital and operating costs. Its downside is the loss of operational flexibility, as the reaction conditions may not be compatible with or optimal for membrane operation. MR applications (e.g., homogeneous catalytic applications) also exist in which the distinction in configuration is not clear.

There is a multitude of proposed configurations which combine the membrane and reactor into a single unit. For catalytic applications, Sanchez and Tsotsis [1] identified six such configurations, which are listed in Table 1 and shown schematically in Figure 2. The most common configuration is the PBMR, in which the membrane provides only the separation function. In catalytic applications, the reaction function is provided by a packed bed of catalyst particles placed in the interior or exterior reactor compartments. In the CMR configuration the membrane is itself catalytic, thus providing simultaneously the separation and reaction functions. To accomplish this for biological reactions, for example, enzymes or bacteria are grafted onto the membrane. Sometimes, the membrane used in the PBMR configuration is also itself catalytically active, often unintentionally so (frequently the case for metallic membranes), but, on occasion, purposely so to pro-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.d16_d01

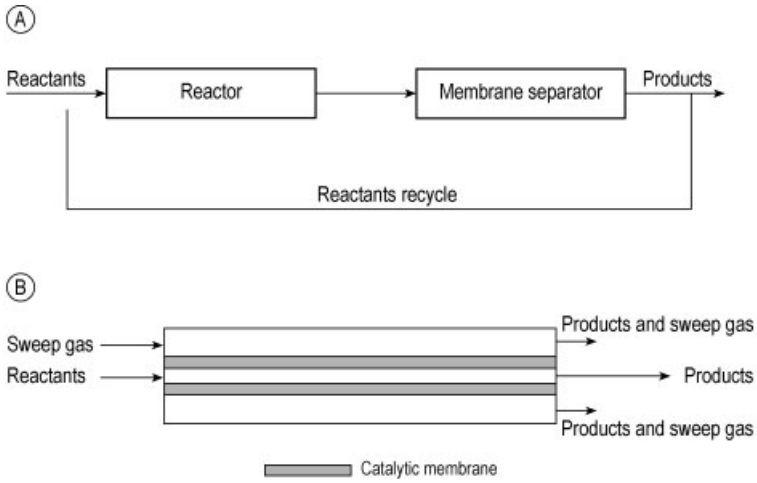

Figure 1. Different membrane reactor process configurations.

vide an additional catalytic function. The membrane reactor in this case is named PBCMR. For better control of the process temperature (important in the case of exothermic reactions), replacement of the packed bed by a fluidized bed (FBMR or FBCMR) has been suggested. In the CNMR configuration, the membrane is not typically permselective. It is only used to provide a reactive interface or as a contactor. Here the discussion is expanded to cover two other classes of membrane reactors. These are membrane bioreactors (MBR), in which the reaction is catalyzed by enzymes or whole-cell microorganisms, and pervaporation membrane reactors (PVMR), in which the separation task is carried out by a pervaporation membrane. For these membrane reactors both types of process configurations shown in Figure 1 are currently in use. For the process design which combines the membrane and reactor into a single unit, one could potentially identify a number of different configurations in analogy to those listed in Table 1 for the CMR.

Table 1. Classification of membrane reactors [1]

Abbreviation	Description
CMR	catalytic membrane reactor
CNMR	catalytic nonpermselective membrane reactor
PBMR	packed-bed membrane reactor
PBCMR	packed-bed catalytic membrane reactor
FBMR	fluidized-bed membrane reactor
FBCMR	fluidized-bed catalytic membrane reactor

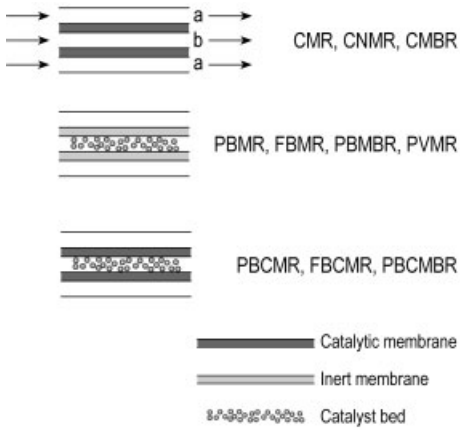

Figure 2. Different membrane reactor configurations a) Shell side; b) Tube side

Though the first publications on MRs appeared in the late 1960s [2], [3], most of the progress in the area was made in the 1980s and 1990s [4], mirroring the state of the membrane field as a whole. In the field of catalysis, in particular, there has been a renewed interest in MRs in the 1990s, motivated by the development of stable microporous/dense inorganic membranes. These offer advantages over organic membranes, because they are stable at higher temperatures (> 373 K) and have good chemical and mechanical resistance [5], thus broadening the range of operating conditions for MR applications. This significant upturn in activity is reflected in the increasing number of

publications on catalytic MRs. Detailed descriptions have been published along the way. For recent accounts of the state of the art, see [6–8].

Both dense and porous membranes have found use in MR applications. Dense membranes are made of polymers, metals (mostly noble metals like Pd, Pt, Ru, Rh, Ir, and Ag) and their alloys, and solid oxides. Dense metal membranes (particularly Pd) were used in early applications [9]. They were used in the former Soviet Union by Gryaznov et al. [10–13]. They have high selectivity towards hydrogen or oxygen (Ag), but their high cost, relatively low permeability, and limited availability have hindered their extensive industrial application. There is also concern with their mechanical properties (embrittlement and metal fatigue) for applications requiring repeated thermal cycling at high temperatures. Ongoing efforts for improving the characteristics of these membranes [7], [8] involve depositing thin metallic films on underlying porous substrates with the aim of improving membrane permeance and cost without unduly impacting selectivity. Another type of dense membranes attracting attention in recent years [14] consist of solid oxides (ZrO_2, Y_2O_3, Bi_2O_3) and solutions of mixed solid oxides (perovskites, brownmillerites, etc.). These materials act as solid electrolytes that allow the transport of oxygen or hydrogen. They are finding use in catalytic MR applications involving partial and total oxidation reactions.

Porous membranes composed of glass, ceramics, or polymers, have also found use in MR applications. They are classified (IUPAC) as macroporous (average pore diameter $d_p > 50$ nm), mesoporous ($50 < d_p > 2$ nm), or microporous ($d_p < 2$ nm). Earlier studies made use of commercially available mesoporous membranes (Vycor glass; alumina, titania, or zirconia [15]). Recent efforts have focused on increasing the selectivity of these membranes towards small molecules. This has been accomplished by decreasing the mean pore size and/or narrowing the pore-size distribution of existing mesoporous membranes [8] or by synthesizing new microporous membranes (e.g., zeolite membranes [16]). Porous polymeric membranes have been the design choice in membrane bioreactors (MBRs); they have also been recently used for low-temperature catalytic applications [17].

2. Catalytic Membrane Reactors

2.1. Dehydrogenation Reactions

The catalytic dehydrogenation of light alkanes is a potentially important process for the production of alkenes, which are valuable starting materials for a variety of applications. The reaction is endothermic and is performed at high temperatures to improve the alkene yield, which is limited at lower temperatures by the thermodynamic equilibrium. Operation at high temperatures, however, results in catalyst deactivation, which requires frequent regeneration, and the production of undesired byproducts. Because of the promise of CMRs to overcome equilibrium limitations, this reaction was an obvious choice for MR applications from the beginning, and still remains the most commonly studied reaction. MR studies of light alkane dehydrogenation reactions to produce olefins have made use of both dense and porous inorganic membranes. Dense metallic membranes proved particularly succesful in the laboratory [18] because of their very high selectivity towards hydrogen, which eliminates reactant loss, a phenomenon that adversely impacts alkene yield for porous membranes. Their industrial applications have been hindered by problems of mechanical stability, high cost, limited commercial availability, poisoning by sulfur/chlorine impurities, and deactivation by carbon deposition. Some of the key studies are summarized in Table 2.

2.2. Hydrogenation Reactions

Catalytic MRs have also been applied to catalytic hydrogenation reactions, where the membrane's role is to separate the liquid from the gaseous reactant (e.g., hydrogen) and to provide a means for delivering this reactant at a controlled rate. Industrial applications of the technology have been reported for dense metallic membranes. They include the synthesis of vitamin K from quinone and acetic anhydride [34], and the hydrogenation of *cis*/*trans*-2-butene-1,4 diol to *cis*/*trans*-butanediol [35]. To provide a source of hydrogen, Gryaznov et al. proposed coupling through the membrane a catalytic dehydrogenation with a hydrogenation reaction.

Table 2. Membrane reactor applications for hydrocarbon dehydrogenation reactions

Hydrocarbon	Reactor type and membrane	Main results	Refs.
Ethane	600 – 873 K; Pt/alumina membrane or catalyst; composite membrane (PdAg alloy/glass); PBMR, PBCMR, CMR	20 % conversion increase over equilibrium	[19–21]
Propane	673 – 900 K, Pt or Cr_2O_3/alumina catalyst; PdAg, dense silica, or silica/alumina porous membranes; PBMR	conversion up to 3 times higher than equilibrium	[18], [22–25]
Isobutane or butane	723 – 823 K, PtSn or Cr_2O_3/alumina catalyst; dense silica, alumina, zeolite or composite (Pt/alumina) membrane; PBMR	isobutene yield up to 6 times higher than equilibrium	[26–30]
Cyclohexane	Pd metal, carbon hollow fiber; PBMR	conversions higher than equilibrium	
Ethylbenzene	800 – 973 K, Fe_2O_3 catalyst; alumina or composite (Pt/alumina, Pt/SS) membrane; PBMR	styrene yield up to 1.5 times higher than equilibrium; better selectivity	[31–34]

In this concept the dehydrogenation reaction on one side of the membrane produces hydrogen, which diffuses through the membrane and participates in the hydrogenation reaction [10], [12], [36]. This approach also provides a means for coupling the reactions energetically. Other novel concepts for providing hydrogen include integrating the hydrogenation reaction with H_2O electrolysis [37], [38]. In addition to dense metal membranes a number of recent studies also report the use of proton-conducting solid oxide membranes [39], [40] for hydrogenation reactions. Porous membranes have also been used for this type of reaction. For the reported applications the membrane acts both as a catalyst and as a contactor between the reactive phases. An advantage of the use of MRs for such an application are the decreased mass transfer limitations compared with the more traditional stirred-tank slurry reactor. Some of the key studies are summarized in Table 3.

2.3. Oxidation Reactions

Selective catalytic hydrocarbon oxidation is difficult to implement because the desired products are generally more reactive towards O_2 than the original hydrocarbons. One way to increase product selectivity is to control the O_2 concentration along the reactor length. This can be implemented by using an MR. The membrane allows the O_2 and the hydrocarbon to be fed in separate reactor compartments. With the membrane reactor the O_2 concentration along the reactor length can be controlled to influence selectivity, something not possible for fixed-bed reactors, for which the O_2 concentration is maximum at the entrance and decreases monotonically along the length. Therefore, for such reactors, the selectivity is low at the reactor inlet, where the reaction rate is the highest, and this adversely impacts overall yield. One potential additional benefit of membrane reactors for catalytic partial oxidation is that they provide for conditions less prone to explosions and other hazards typically associated with the oxidation of hydrocarbons [7].

The studies so far make use of both dense and porous membranes. Dense membranes consist of metallic silver (unique among metals in that it permeates oxygen) and its alloys, but most of the recent efforts make use of solid oxide membranes. The earlier such membranes were solid oxide electrolytes, like titania, yttria, calcia, or magnesia-stabilized zirconias (TiYSZ, YSZ, CSZ, MSZ) [45], [46]. These materials have good mobility of oxygen in the lattice but lack sufficient electronic conductivity, and this necessitates the use of an external circuit. Significant advances in this area have been made by the introduction of perovskites [47] and brownmillerites [48], [49], which have both good ionic and electronic conductivity and are also more thermally stable.

Oxidative methane coupling to produce ethylene and ethane has been investigated in solid-oxide MRs. Membranes used include TiYSZ, YSZ, CSZ, MSZ, doped Bi_2O_3, and oxygen- and proton-conducting perovskites with or with-

Table 3. Membrane Reactor Applications for hydrogenation reactions

Reaction	Operating conditions	Main results	Refs.
Vitamin K synthesis	400 K, Pd, Ni (tubular) membrane, CMR	yields up to 95 % in one step instead of 80 % in four steps	[34]
Cis/*trans*-2-butene-1,4 diol to *cis*/*trans*-butanediol	330 – 360 KPd, Ni or Pd – Ru (foil) membranes, CMR	up to 96 % yield of *cis*-butanediol	[35]
Cyclopentadiene hydrogenation	380 K, Pd – Ru composite ceramic – metal and metallic membrane, CMR	yields up to 1.3 times that of a batch reactor	[41]
α-methylstyreneto cumene	320 K, Pd/alumina membrane, CMR	cocurrent catalytic membrane reactor more efficient than a trickle-bed reactor	[42]
Nitrobenzene to aniline	320 K, Pt/alumina membrane, CMR	reactor effectiveness is strongly influenced by the configuration	[43]
Cinnamaldehyde hydrogenation	300 K, Pt/alumina membrane, CMR	modest increases in hydrocinnamyl alcohol production compared to batch slurry reactor (up to 1.1 times)	[44]

out the application of an external potential [6]. Though simulations [50] predict yields as high as 50 %, reported experimental yields so far, are less than 30 %, which is also the maximum yield obtained with packed-bed reactors [51]. The selectivity in the MR is reported to be higher than that in conventional reactors [52], [53]. Significant additional yield improvements would result [54] if the membranes could also simultaneously remove the C_2 product, but such membranes are currently not available [7]. Dense membranes have good oxygen permselectivity, which allows the use of air for oxidative coupling. However, their permeability is low, which limits reactor efficiency. This has motivated the use of porous membranes. Recent investigations using the PBMR configuration are encouraging. Using nonpermselective, commercially available alumina membranes and conventional oxidative coupling catalysts gave promising yields (ca. 25 %) [55–58]. An La-stabilized porous γ-alumina membrane in a PBMR packed with $Mn-W-Na/SiO_2$catalyst was reported to give higher C_2 yields [59] than those obtained with the more conventional co-feed configuration.

Another partial oxidation reaction attracting attention is synthesis gas production from methane. The earlier efforts made use of solid electrolytes [60], [61]. Integration of CH_4 partial oxidation with steam reforming in an electrochemical MR gave high syngas yields (97 %) [62]; the overall system is potentially thermoneutral. More recent investigations involve internally shorted metal/YSZ composites or single-phase mixed conductors [63], [64], which need no external circuitry. These materials have been shown to be stable after more than 500 h of synthesis gas generation. Collaborations by Amoco and the Argonne National Laboratory [65–69] used perovskite-like $Sr(Co,Fe)O_x$ materials and achieved conversion efficiencies of greater than 99 % and stable operation for over 900 h. Brownmillerite membranes [48], [49] also gave high yields and stable operation for over one year. Two U.S. industrial consortia were recently formed [8], [70] to commercialize solid-oxide MR-based synthesis gas production. One consortium, supported in part by the U.S. DOE., is led by Air Products and Chemicals, and includes Babcock and Wilcox, Ceramatec, Eltron, ARCO, Penn State University, and the University of Pennsylvania. This group has announced plans for an eight-year, $ 85 x 10^6 programme. The other group is an industrial alliance among BP/Amoco, Praxair, Statoil, SASOL, and Phillips Petroleum, also enlisting the help of a number of Universities including MIT, and the Universities of Alaska Fairbanks, Houston, Missouri-Rolla, and Illinois-Chicago. A third consortium with the same purpose seems to be in planning in Europe [8]. Despite these successes, concerns still remain about long-term materials stability in the high-temperature reactive environments [71]. Efforts to produce syn-

thesis gas by using a nonpermselective porous alumina membrane are also under way [72].

Other partial hydrocarbon oxidation reactions studied in membrane reactors include methanol and formaldehyde production from methane; the oxidative dehydrogenation of low molecular mass hydrocarbons, including ethane to ethylene, ethylene to acetylene, propane to propylene, and butane to butene and maleic anhydride; and selective oxidation of $C_2 - C_4$ alkanes into various oxygenates. A recent application involves the use of polymeric membranes in which active catalysts are incorporated as interfacial contactors to carry out partial oxidation (but also hydrogenation and hydroformylation) of various organic compounds [73], [74]. This concept allows the reaction to be carried out without a cosolvent, and prevents loss of catalyst. This was demonstrated, for example, for the oxyfunctionalization of *n*-hexane by H_2O_2 to a mixture of hexanols and hexanones by using titanium silicalite catalyst imbedded in a modified PDMS membrane [75]. Some key studies are listed in Table 4.

2.4. Other High-Temperature Applications

MRs have also been used for reactions of relevance to environmental protection. Such reactions include H_2S decomposition [110], [111]; NH_3 decomposition [112], [113], of potential interest in the treatment of coal gasification streams; and hydrodechlorination of dichloroethane [114]. For a number of applications in which the goal is full conversion of the pollutant and avoiding slippage out of the reactor, the membrane acts as a reactive interface between reactants flowing on either side. The concept was first demonstrated [115] for the selective catalytic reduction (SCR) of nitric oxide with ammonia. The membrane provided an interface for the nitric oxide to completely react with ammonia. The Claus reaction between hydrogen sulfide and sulfur dioxide [116], [117] and the oxidation of CO [118] and H_2S [119] were also studied in a CMR with a nonpermselective membrane. For these reactions, when the rate is fast, a sharp reaction front forms within the membrane and prevents reactant slippage from either side. Oxidative destruction of VOCs r [120–122] and photocatalytic oxidation of carbonyl compounds and bioaerosols [123] have been studied in a CMR. In this case, effective contact of the VOC with the catalyst is attained by forcing the VOC-containing gaseous stream through the membrane.

Steam and dry (i.e., CO_2) reforming of methane is one of the most important reactions for which membrane (porous and dense) reactors have been applied. Membranes prepared by coating thin Pd films onto mesoporous substrates by an electroless plating technique are currently being evaluated by Tokyo Gas and Mitsubishi Heavy Industries in membrane reformers in polymer electrolyte (PEM) fuel cell systems [124]. The same reaction was studied in a pilot plant scale FBMR system (Fig. 3) [125–127] which combines a fluidized bed of $Ni - Al_2O_3$ catalyst with a bundle of twelve tubular Pd membranes. Reactor performance was shown to be sensitive to the way in which the membrane area is distributed between the dense and dilute fluidized-bed phases. Methane and steam conversion enhancements were observed of up to 1.3 times that of the same reactor without a membrane. Despite the high installed membrane permeation capacity, the low permeability of Pd membranes still limited the conversion enhancement. Though membrane reformers have received significant attention, they have yet to be commercialized. Questions still remain regarding the effect that hydrogen removal may have on catalyst coking and H_2S poisoning of the membranes [128]. The overall economics still remain debatable, at least in the context of large-scale applications [129], [130]. Methanol membrane reformers for H_2 production have been studied [131], [132] in the context of mobile fuel cell applications. Potential benefits include reduced reactor volumes and a more compact unit design.

MRs combining the water gas shift (WGS) reaction and hydrogen separation in a single unit (WGSMR) are attracting attention in the context of the Integrated Gasification Combined Cycle (IGCC) plant, as a means to reduce CO_2 emissions. A study by a European academic and industrial consortium [133] concluded that the net efficiency of the WGSMR-integrated IGCC plant is 42.8 % (lower heating value = LHV) with 80 % CO_2 recovery, as opposed to 46.7 % (LHV) for an IGCC plant

Table 4. Membrane reactor applications for hydrocarbon oxidation reactions

Hydrocarbon and reaction	Operating conditions	Main results	Refs.
Methane oxidative coupling	1023 – 1123 K	Conversion up to 30 %	[53], [55–59], [76–88]
	YSZ, CSZ TiSZ, MSZ, Bi_2O_3, perovskites, Al_2O_3 or La_2O_3/Al_2O_3, and Pd – Ag membranes conventional oxidative coupling catalysts, Mn – W – Na/SiO_2 sometimes external electrical potential applied; CMR, PBMR, 2-stage PBMR	selectivity up to 97 %	
Methane partial oxidation to syngas	873 – 1273 K; CSZ, YSZ, perovskites, brownmillerite membranes; PBMR	conversion up to 99 %; selectivity to H_2 and CO up to 97 %	[48], [49], [61,62, 64–69], [89], [90]
Methane partial oxidation	573 K, zirconia – alumina or zirconia – calcia – alumina membrane; CMR	selectivity up to 96 %, generally low conversions	
to formaldehyde			[91], [92]
to methanol			[93], [94]
Ethane, oxidative dehydrogenation	700 – 900 K; TiYSZ, YSZ, Ce membranes	yields up to 7.5 %,	[77], [95], [96]
partial oxidation to oxirane and other oxygenates	Ag or V_20_5 catalyst; CMR, PBMR	conversion up to 6 times higher than equilibrium	[97], [98]
Ethylene partial oxidation to oxirane and other oxygenates	V_20_5/AL_2O_3; CMR, PBMR	higher yields than conventional reactors; unique selectivity behavior necessitates C_2H_4 distributed feed.	[80], [99], [100]
Propane, partial oxidation to propylene oxide	800 – 900 K; TiYSZ, YSZ, Ce, V_20_5/Al_2O_3 membranes	yields up to 7.5 % in epoxide	[22], [77], [97], [98]
oxidative dehydrogenation	Ag or V_20_5 catalyst; CMR or PBMR	yields up to 52 % in propene (1.4 times equilibrium)	[101–105]
Isobutane oxidative dehydrogenation	673 – 973 K, Pd membranes, Pt/Al_2O_3 catalyst	potential for autothermal operation	[107]
Butane, oxidative dehydrogenation	700 – 800 K; membranes: TiYSZ, YSZ, V_20_5/Al_2O_3	propene yields 1.7 times equilibrium	[77], [106]
partial oxidation to oxygenates including maleic anhydride	Ag or V_20_5 or VPO/catalyst; CMR or PBMR	anhydride yields up to 17.7 %	[97], [98], [108], [109]
Cyclohexane and dodecane partial oxidation to alcohol and ketones	298 K; PDMS membrane; zeolite-encapsulated iron phthalocyanine incorporated in membrane; CMR	higher turnover rates and selectivity with respect to homogeneous reactor	[73]
Cyclic alcohols (C_1, C_6, C_7) and 3-penten-2-ol epoxidation	298 K; PDMS membrane; zeolite-encapsulated iron-phtalocyanine and TDCPP(Mn)Cl complexes incorporated in membrane; CMR	higher turnover rates (9 ×) and selectivity (5 ×) than homogeneous reactor	[74]

without CO_2 recovery, and 40.5 % (LHV) with conventional CO_2 removal. Moreover, an economic analysis indicates favorable investment and operational costs. The study concluded that the WGSMR-integrated IGCC process is technically feasible, but further development in the area of high-temperature hydrogen separation membranes remains essential. Another application of the WGSMR is tritium recovery from breeder-blanket fluids in fusion reactor systems. The proposed process [134] uses two membrane reactors, the first of which removes the hydrogen isotopes from the purge gas (He) by oxidation, and the second is a WGSMR recovering tritium

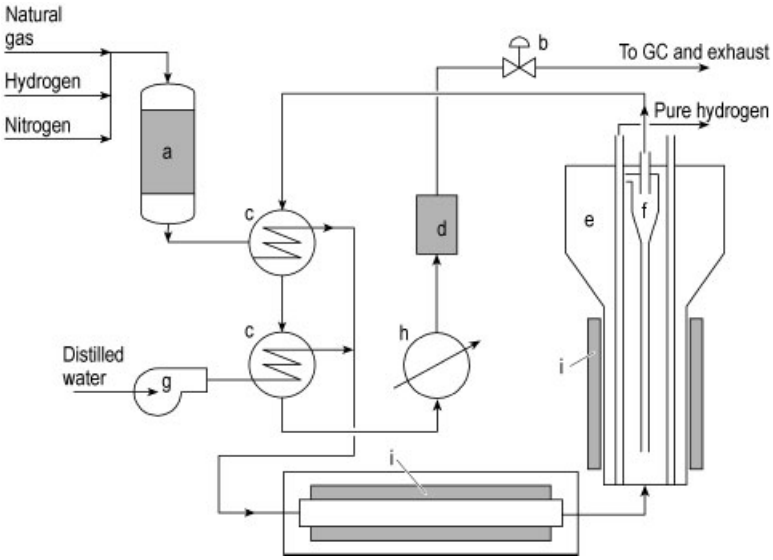

Figure 3. Schematic of the pilot plant for methane steam reforming
The reforming reactor contains a bundle of twelve Pd tubular membranes and a fluidized bed of catalysts
a) Desulfurizer; b) Back-pressure regulator; c) Heat exchangers; d) In-line filter; e) Reformer; f) Internal cyclone; g) Water pump; h) Condenser; i) Heaters
Reproduced from [126] with permission

Table 5. Other high-temperature applications of membrane reactors

Reaction	Operating conditions	Main results	Refs.
H_2S decomposition	773 – 1200 K; porous Vycor glass/Pt – SiO_2 – V composite dense membranes or ZrO_2 – SiO_2 membranes; sulfided Mo catalyst; PBMR	conversions up 3 × equilibrium at 1078 K (feed: 0.84 % H_2S in CH_4)	[110], [111], [135]
SCR of NO_x with NH_3	470 – 620 K; alumina and titania membranes impregnated with V_2O_5; catalytically impregnated fly-ash filter; CMR	Conversions of NO_x up to 70 %; selectivity to N_2 up to 75 %; simultaneous particle removal	[8], [115], [136]
NH_3 decomposition	700 – 900 K; Pd alloy or composite palladium – alumina membranes; PBMR	Conversion up 1.8 × equilibrium	[112], [113]
VOC oxidation	540 – 690 K; membranes: perovskite type (La-X-Y) on alumina tube or Al_2O_3; CMR	conversion up to 100 % , 1.7 times that of the monolith configuration	[120–122], [137]
Claus reaction between H_2S and SO_2	Al_2O_3, CMR	prevents slippage of reactants	[116], [117]
H_2S and CO oxidation	Al_2O_3, CMR	prevents slippage of reactants	[118], [119]
CH_4, steam reforming dry (CO_2) reforming	720 – 1000 K membranes: Pd or Pd/Vycor glass; Al_2O_3; CVD Pt, Ru/Al_2O_3. CMR, PBMR, FBMR	conversions up to 1.2 times equilibrium	[125–127], [138–143], [144], [145]

from tritiated water. Table 5 summarizes some key studies in this area.

3. Pervaporation Membrane Reactors

Pervaporation membrane reactors (PVMR) have attracted attention recently [146], [147]. Pervaporation (PV) has established itself as a practical technology in applications such as the dehydration of organic compounds, their removal/recovery from aqueous solutions, and the separation of organic mixtures. During PVMR operation the membrane removes either the desired (biotechnological/wastewater applications) or the undesired product (e.g., water from esterification reactions). Existing or proposed near-term industrial processes for application of PVMR include the production of ethyl and butyl acetate, ethyl and *n*-butyl oleate, diethyl tartrate, dimethyl urea, ethyl valerate, isopropyl and propyl propionate, and methyl isobutyl ketone.

The principal reactions studied by PVMR are esterifications. In the most common PVMR configuration the membrane and the reactor are housed in separate units (see Table 6). A number of commercial units are in operation [146], [147]. The process layout of a medium-sized esterification plant employing a PVMR, as described, for example, in [148], consists of a cascade of two reactors, each individually coupled to two PV units. Some earlier studies incorporated the membrane and reactor in the same unit [149]. Such a PVMR [150] was used to study the production of butyl oleate by lipase-catalyzed esterification with a hydrophilic cellulose acetate membrane. Conversion in the PVMR was 92 % compared to 61.1 % without PV. The economics of a number of process layouts involving PVMR have been discussed [151]. PVMR studies also exist for reactions other than esterification, including the production of dimethyl urea, methyl isobutyl ketone, and MTBE from methanol and *tert*-butyl alcohol (Table 6).

4. Membrane Bioreactors

Membrane bioreactors (MBRs) are a natural outgrowth of membrane applications in the food industry [160], which include the concentration of proteins from milk and the separation of whey fractions (e.g., lactose, proteins, minerals, and various fats). MBRs are finding increased use in bioengineering. Since the first MBR study on coupling UF membranes with a bioreactor [161], many other applications have been developed. MBRs are used today, for example, during biochemical synthesis to remove reaction products inhibiting biological function or to separate costly biocatalysts from the product stream for recycle (see Fig. 4). This increases cellular concentration, prolongs biocatalyst life, and increases the product turnover rate. Other MBR applications involve the hydrolysis of macromolecules (e.g., proteins), and enzymatic reactions which require co-factor recycling and the efficient delivery of one of the reactants.

In the simplest MBR applications, the membrane separates the products or metabolites of the reaction while retaining the biocatalysts, which are then recycled to the reactor. For whole-cell processes, microfiltration (MF) or ultrafiltration (UF) membranes are generally used to separate the biomass (average pore size 0.1 to several micrometers) from the product molecules. In enzymatic transformations UF membranes, with cut-offs ranging from 5000 to 100 000 Da, are utilized to retain the enzymes, whose molecular mass ranges from 10 000 to 100 000 Da. These simple MBR-based processes have been widely studied in the industry for the production of many chemicals, and efficiency improvements over the conventional bioreactors have been reported. In other MBR applications the biocatalyst is immobilized on the membrane itself. For whole-cell MBR it is fixed within the membrane's porous structure. For enzymatic MBR the biocatalyst is coupled to the membrane either by grafting (covalent or ionic binding or by cross-linking), gelation, physical adsorption, or by entrapment between the two skin layers of a porous membrane [162].

4.1. Whole-Cell MBR Applications

Coupling membrane separation with whole-cell (plant or mammalian cells, bacteria, or yeast) biocatalysis (fermentation, oxidation, etc.) is a common application of MBR. The production of lactic and propionic acid in a whole-

Table 6. Applications of pervaporation membrane reactors

Product	Operating conditions	Main results	Refs.
Butyl acetate	anion-exchange resin catalyst ;reactor and membrane separate units	PV replaces distillation; energy savings	[152]
Enzymatic esterification oferucic acid and hexadecyl alcohol	hollow-fiber membranes; reactor and membrane separate units	MR attains over 90 % conversion (vs. 53 % for conventional reactor)	[153]
Production of methyl isobutyl ketone	hydrophilic polymeric membrane; two reactors with the membrane unit in between	enhanced yield	[155], [156]
Butanol production	silicone membrane	improvements of 200 % in productivity	[154]
Dimethylurea production	PVA hydrophilic membrane; reactor and membrane separate units	removal of water allows recycle of CO_2 and amine resulting in reduced energy costs and by-products	[155]
Coating resins by polycondensation	Silica microporous membranes; reactor and membrane separate units	improved yield; membrane shows high water permeance and permselectivity	[157]
Esterification reactions of carboxylic acids and ethanol	hydrophilic polymeric membranes (PVA, polyimide, polyetherimide); reactor and membrane in same unit	significantly improved yields	[149–151], [158], [159]

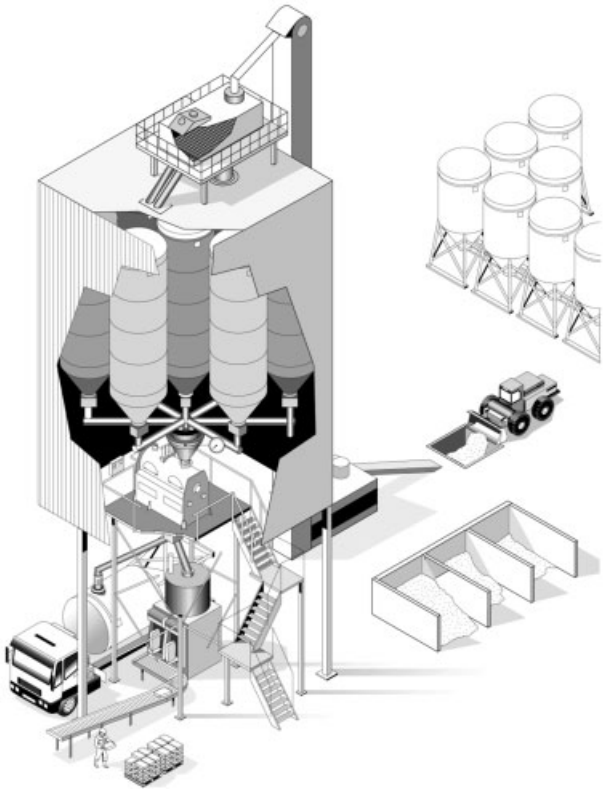

Figure 4. Bioreactor coupled with an ultrafiltration (UF) unit

cell MBR using two different types of bacteria has been reported [163], [164]. Ethanol production was studied with *Saccharomyces cerevisae* [165]. Whole-cell MBRs have been utilized in a number of biochemical syntheses. An example used industrially is biosynthesis of growth hormone by the bacterium *E. coli* [166]. Other industrial examples include the synthesis of homochiral cyanohydrins [167], the production of L-aspartic acid [168], [169]. Naphthalene was oxidized to optically pure *cis*-1,2-dihydroxy-1,2-dihydronaphtalene by *Pseudomonas fluoroscens* [170]. Phenylacetylcarbinol, an essential intermediate for the production of the nasal decongestant ephedrine, was synthesized by using *Candida utilis* [171]. Whole-cell MBRs have also been applied to two-phase biotransformations, which are employed when one or more reactants and/or products have low solubility in water. In this case the membrane has a double function: it separates the two phases, thus avoiding the formation of an emulsion phase, and extracts the reactants from the organic phase and brings them into contact with the aqueous reaction phase. This type of reactor was used, for example, for the reduction of geraniol to citronellol (both poorly water-soluble) by *Saccharomyces cerevisiae* and the epoxidation of 1,7-octadiene by *P. oleovorans* [172].

In the perfusion hollow-fiber MBR, used for the synthesis of monoclonal antibodies, the cells grow in the extracapillary space (i.e., between the fibers). Nutrients circulating in the fibers themselves diffuse through the thin fiber skin to nourish the cells, while the metabolites are simultaneously extracted through the same hollow fiber skins [173]. Perfusion hollow-fiber MBRs mimic the function of the organs of living organisms, which also make use of a system of small capillaries. For mammalian cells (also called hybridomas) this type of MBR has been shown to support cell concentrations as high as 10^8 cells per milliliter [174]. These high cell concentrations result in highly concentrated antibody and protein harvests, with antibody harvests as high as 17 mg/mL [175], [176]. Such high yields and the relatively simple and inexpensive operation of hollow-fiber MBRs has made them an attractive choice in a number of commercial applications for production of monoclonal antibodies [173], [177]. Hollow-fiber MBRs have also been used for the production of a number of other fine chemicals, for example, the production of isovaleraldehyde from isoamyl alcohol using *Gluconobacter oxidans* [178]. Some of the most important whole-cell MBR applications are shown in Table 7.

4.2. Enzymatic MBRs

There are many important enzymatic reactions for which MBRs are finding application. In MBRs enzymes are often immobilized on the surface of the membrane or in its porous structure. Free enzymes and enzymes grafted onto a soluble polymer to increase their molecular size are utilized. In the latter case the membrane simply acts as a barrier to retain the enzymes, while allowing the removal of products and/or the addition of reactants. A key advantage of immobilization is that it permits easier enzyme recovery; in addition, grafting the biocatalyst sometimes increases its thermal stability and its resistance towards organic solvents.

MBRs are widely used in the enzymatic hydrolysis of macromolecules [165], [183]. They have been applied, for example, to protein hydrolysis and to enzymatic reactions requiring co-factor recycling [184], [185]. For the latter class of reactions, one of the earliest successes was L-phenylalanine synthesis from acetamidocinnamic acid utilizing NADH (β-nicotinamide adenine dinucleotide) as a co-factor [186]. Coenzymes like NADH act as co-substrates during the enzymatic reaction and, in so doing, they are converted into an inactive form. Since they are expensive ($ 1000/mol) they must be recovered for regeneration. In the above study [186], the co-factor was grafted onto a soluble polymer and separated by an UF membrane. Several studies involving this type of enzymatic transformation in MBRs have been reported [187]. L-*tert*-Leucine [188], erythrulose, [189] and (*S*)-2-octanol [190] have been synthesized in various MBR configurations.

Hydrolysis or transesterification reactions of lipids catalyzed by lipase and phospholipase are another reaction class to which MBRs have been applied. The enzymatic hydrolysis of olive oil by lipase was studied in a MBR using a hydrophobic membrane [191]. Lipase-catalyzed hydrolysis of olive and sunflower oils was also studied [192], [193]. MBRs have also found use in the

Table 7. Applications of whole-cell membrane reactors for biosynthesis

Reaction	Biocatalyst and type of MBR	Main results	Refs.
Alcoholic fermentation	*S. cerevisae*; inert membrane	conversion increases up to 3000 % with respect to a batch reactor	[114]
Growth hormone synthesis (hGH)	*E. coli*; inert membrane	yields by batch equivalent up to 40 000 extracted hypophyisis; hormone free of pathogens (e.g., Creutzfeld – Jacob syndrome)	[166], [179]
Propionic and lactic acid production	*P. bacterium*, *L. delbruecki*, inert membrane	conversion enhancement up to 1.6 times with respect to the stirred-tank reactor	[164]
β-amilase production	*Clostridium thermosulfurogenes*; inert membrane	yield up to 3000 % with respect to a batch reactor	[180]
L-Aspartic acid production	*E. coli*; inert membrane	maximum rate of production of 4.5 mol L^{-1} h^{-1}; industrial-scale process, continuous production for up to 30 d	[168], [169]
Isovaleraldehyde production	*G. oxidans*; inert membrane	conversion up to 90 %	[178]
Naphthalene oxidation to *cis*-1,2-dihydroxy-1,2-dihydronaphtalene	*P. fluoroscens*; inert membrane	production rate increases up to 300 % with respect to a batch reactor	[170]
Phenylacetylcarbinol synthesis	*C. utilis*; inert membrane	higher productivity and cell concentration (+ 300 %) with respect to a batch reactor	[171]
Antibodies synthesis	mammalian cells (hybridoma) immobilized on the membrane; also with an inert membrane	high cell concentration, good antibody productivity and purity; production up to 30 d continuously	[173], [177], [181]
Rifamycine B	*Nocardia mediterranei*, immobilized on the membrane	yield up to 1500 % with respect to a batch reactor.	[182]
Lactic acid	*L. rahmnosus*, immobilized on the membrane	conversion enhancement up to 1.6 times with respect to a stirred-tank reactor	[163]

production of polysaccharides. MBRs provide a simplified method for producing these materials, which are, typically, obtained in small quantities from natural sources by elaborate procedures. The synthesis of inulin (poly-β-(2,1)-fructose) in a MBR with the aid of fructosyltransferase immobilized in the porous structure of the membrane [194].

For enzymatic conversions, in particular, hollow-fiber MBRs are widely used, with enzymes typically immobilized in the porous structure of the hollow fibers. An advantage of these MBRs over the conventional bioreactors is the longer contact times of reactants with the enzymes, due to their high surface-to-volume ratios. Low residence times are a problem with conventional enzymatic bioreactors, and good efficiency is obtained only with rapid reactions. An example of the use of such a hollow-fiber MBR is the production of (2*R*,3*S*)-methyl methoxyphenylglycidate ester (2*R*,3*S*)-methyl methoxyphenylglycidate ester [(2*R*,3*S*)-MMPG], a chiral intermediate for diltiazem, an important calcium channel blocker used for the treatment of hypertension and angina. This MBR process was developed by Sepracor and was applied on an industrial scale by Tanabe Corporation in Japan. The reaction consists of the racemic resolution of the (±)-methyl methoxyphenylglycidate ester [(2*R*,3*S*)-MMPG + (2*S*,3*R*)-MMPG] by lipase, which is entrapped within the porous structure of hollow-fiber membranes, which separate an organic and an aqueous solution. Methoxyphenylacetaldehyde, which is an enzyme inhibitor, is also produced as a byproduct, but it is eliminated by adduct formation with bisulfite present in the aqueous phase [195], [196]. A schematic of the hollow-fiber MBR is presented in Figure 5. Some of the most important MBR applications are listed in Table 8.

A drawback of MBRs, especially when whole cells or bacteria are immobilized in the fiber pores, is biofouling, typically manifested by dra-

Table 8. Applications of enzymatic membrane reactors for biosynthesis

Reaction	Enzyme and type of MBR	Main results	Refs.
l-Leucine synthesis	l-leucine dehydrogenase and an enzymatic co-factor: NAD (grafted onto a polymer); inert membrane	high productivity (100 %) over a 7-week period	[197]
l-phenylalanine synthesis	acylase/L-PHEDH/FDH∗ and an enzymatic co-factor: NADH (grafted onto a polymer)	low deactivation rates of enzymes, and substrate conversion up to 90 %; continuous production for up to 60 d	[186]
Amino acids, enzymatic racemic resolution	acylase	15 – 20 t per month of l-amino acids	[198]
(2*R*,3*S*)-methyl methoxyphenylglycidate ester synthesis	Lipase imobilized on a membrane	commercial process (Sepracor, Tanabe); 75 t/a production	[195], [196]
Inulin (poly-β-(2,1)-fructose) synthesis	fructosyltransferase imobilized on a membrane	low-polydispersity product	[194]
Olive and sunflower oil hydrolysis and transesterification	lipase, phospholipase, imobilized on a membrane	high enzyme activity (up to 95 %) under steady state conditions	[189], [192], [193]

∗ (L-PHEDH) = L-phenylalanine dehydrogenase, (FDH) = formate/formate dehydrogenase

matic decreases in the permeate flow leading eventually, if left unchecked, to total plugging. The problem results because of adsorption on the membrane of metabolites that bacteria and cells produce, and of the coagulated proteins from lysed cells. The conventional solution to the problem is to stop the reaction, clean membrane, and sterilize the reactor. Critical to MBR success is the use of membranes and cleaning techniques appropriate for each type of reaction, in order to avoid irreversible plugging. In recent years many imaginative ideas have been proposed for in situ clean-up, and significant improvements in performance have been reported [199].

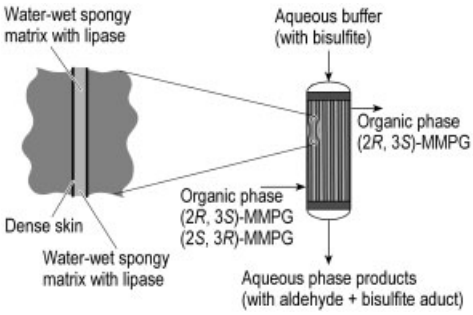

Figure 5. Schematic of the hollow-fiber, two-phase enzymatic bioreactor for the production of (2*R*,3*S*)-methyl methoxyphenylglycidate ester

4.3. Environmental Applications

Another application of MBR is in wastewater treatment, conventionally carried out by biological processes, transforming organic contaminants into harmless products, together with residual sludge. At some point the biocatalyst must be separated from the treated water. Biomass immobilization on porous supports (e.g., biofilters [200]) alleviates the problem somewhat. For heavily polluted waters, however, fast biomass growth clogs the beds and results in frequent shut-downs. MBR-based processes present a solution to the biomass/water separation problem, since membranes can provide a barrier for microbes and other particles, MBRs require a smaller volume than conventional systems, an important consideration in Europe and Japan. They allow for more effective process control, since the residence time in the fermentation vessel and the membrane permeation rate can be independently adjusted [201].

MBRs have been studied (e.g., Lyonnaise des Eaux in France) for the biological denitrification of drinking water and the anaerobic treatment of wastewater from the food industry. Challenges include membrane biofouling, and a large membrane flux to make the process economic, since the price of water is low and the volumes to be treated are enormous. Two notable commercial successful efforts include the Biosep process (OTV Industries, France [202]) and the Kub-

Table 9. Wastewater treatment units in France using the Biosep process (with permission of OTV Industries)

Company	Main activity	DCO treatment capacity, kg/d	Wastewater treatment capacity, m^3/d
Great Lakes Chemicals	Chemicals production	1500	300
Logimer	Food Industry	830	290
Lunor	Food industry	8000	540
Fabergé	Cosmetics	1300	300
Smart	Automobile	300	225
Dicosa (Pierre Fabre)	Cosmetics	2500	510
Faprogi (L'Oréal)	Cosmetics	9000	680
Papeteries du Rhin (Sonoco & Paul)	Paper production	3600	900

ota Process (Kubota Corporation, Japan [203]). Both systems consist of an aerobic bioreactor, with a hollow fiber bundle membrane immersed in the activated sludge, and air blown through the reactor bottom, creating agitation to improve reactor efficiency and diminish biofouling. Using hollow fibers, whose dense skin is on the outside in contact with the suspension, reduces the need for regular membrane back-washing The clean water is extracted by applying vacuum on the inner membrane side. A schematic of the process is shown in Figure 6.

This process has the advantage of generating very low quantity of sludge (with increased concentration and age) when compared with the more traditional biological treatment. Other advantages include its lower operating costs and, due to its compact design, a remarkable reduction in floor space requirements. Many units of different capacity have been installed in France, as indicated in Table 9.

5. Conclusions

Membrane reactors have and can be potentially applied to a broad range of processes, from the catalytic dehydrogenation of hydrocarbons, to hormone synthesis, to wastewater treatment. They are already commercial in low-temperature applications, like the biological production of fine, high-value chemicals, pervaporation, and wastewater treatment. These processes utilize commercially available polymeric and macro-/mesoporous inorganic membranes and have modest requirements in terms of membrane housing and sealing. High-temperature applications are not yet commercial, though good strides have been made in recent years, an example being synthesis gas or hydrogen production by partial oxidation or steam reforming of methane. Further advances here must await the development of more stable and affordable membranes, but the expected benefits are enormous given the size of the large-scale applications in the oil and petrochemical industry. One key hurdle for the application of membrane reactors to these industries is that existing processes are mature and highly optimized, with low a return on investment that does not provide much leeway in terms of initial capital investment. Membrane costs and availability are key factors here. Even for highly value added biological products, membrane capital investment is not negligible. For large-scale petrochemical applications such considerations remain the key factor hindering further progress.

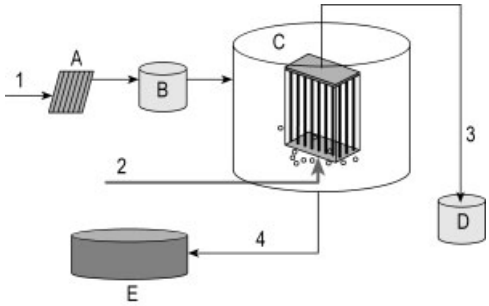

Figure 6. Schematic representation of the Biosep process 1: wastewater; 2: air; 3: treated water; 4: sludge; A: screening unit; B: buffer tank; C: bioreactor with hollow-fiber membrane bundle, D: treated water tank; E: sludge tank and treatment. (with permission of OTV Industries)

6. References

Specific References

1. J. Sanchez, T. T. Tsotsis: "Current Develoments and Future Research in Catalytic Membrane Reactors", in A. J. Burggraaf, L. Cot (eds.): *Fundamentals of Inorganic Membrane Science and Technology,* Elsevier, Amsterdam 1996, chap. 11.
2. B. J. Wood, H. Wise, *J. Catal.* **11** (1968) 30.
3. V. M. Gryaznov, V. S. Smirnov, L. K. Ivanova, A. P. Mishchenko, *Dokl. Akad. Nauk. SSR.* **190** (1970) 144.
4. H. P. Hsieh: *Inorganic Membranes for Separation and Reaction,* Elsevier, Amsterdam 1996.
5. R. Soria, *Catal. Today* **25** (1995) 285.
6. J. Sanchez, T. T. Tsotsis: "Reactive Membrane Processes", in S. Kulprathipanja (ed.): *Reactive Separation Process,* Taylor & Fraucis, New York, 2001.
7. A. G. Dixon: "Innovations in Catalytic Inorganic Membrane Reactors", *Specialist Periodical Reports: Catalysis* **14** (1999) 40.
8. G. Saracco, H. W. J. P. Neomagus, G. F. Versteeg, W. P. M. van Swaaij, *Chem. Eng. Sci.* **54** (1999) 1997.
9. US 3, 290, 4061 996 W C. Pfefferie
10. V. M. Gryaznov, *Platinum Met. Rev.* **30** (1986) 68.
11. V. M. Gryaznov, *Platinum Met. Rev.* **36** (1992) 70.
12. V. M. Gryaznov, V. S. Smirnov, M. G. Slinko in G. C. Bond, P. B. Wells F. C. Tompkins (eds): ″Binary Palladium Alloys as Selective Membrane Catalysts,″ *Proc 6th Intl. Cong. Catal.* **2**, 894, 1976.
13. J. Shu, B. P. A. Grandjean, A. van Neste, S. Kaliaguine, *Can. J. Chem. Eng.* **69** (1991) 1036.
14. H. J. M. Bouwmeester, A. J. Burggraaf in A. J. Burggraaf, L. Cot (eds.): *Fundamentals of Inorganic Membrane Science Technology,* Elsevier, Amsterdam 1996, p. 435
15. V. T. Zaspalis, J. Burggraaf "Inorganic Membrane Reactors to Enhance the Productivity of Processes", in R. R. Bhave (ed.): *Inorganic Membranes: Synthesis, Characteristics and Applications,* Reinhold, New York 19917.
16. J. N. van de Graaf, F. Kapteijn, J. A. Moulijn: "Catalytic Membranes", in A Cybulski J. A. Moulijn (eds.): *Structured Catalysts and Reactors,* Marcel Dekker, New York 1998.
17. D. Fritsch, J. Theis in R. Bredesen (ed.): ″First Results with Inorganic Modified Catalytic Polymer Membranes,″ *Proc. Fourth Workshop: Optimisation of Catalytic Membrane Reactors Systems*, Oslo, Norway, May, 1997, p. 109.
18. H. Weyten, J. Luyten, K. Keizer, L. Willems, R. Leysen, *Catal. Today* **56** (2000) 3.
19. E. Gobina, R. Hughes, *J. Membr. Sci.* **90** (1994) 11.
20. A. M. Champagnie, T. T. Tsotsis, R. G. Minet, I. A. Webster, *Chem. Eng. Sci.* **45** (1990) 2423.
21. A. M. Champagnie, T. T. Tsotsis, R. G. Minet, E. Wagner, *J. Catal.* **134** (1992) 713.
22. UK 2.201.1591988Bitler J. G. A.
23. Z. D. Ziaka, R. G. Minet, T. T. Tsotsis, *AIChE J.* **39** (1993) 526.
24. H. Weyten, K. Keizer, A. Kinoo, J. Luyten R. Leysen, *AIChE J.* **43** (1997) 1819.
25. J. P. Collins, R. W. Schwartz, R. Seghal, T. L. Ward, C. J. Brinker, G. P. Hagen, C. A Udovich, *Ind. Eng. Chem. Res.* **35** (1996) 43.
26. T. Matsuda, I. Koike, N. Kubo, E. Kikuchi, *Appl. Catal. A.* **96** (1993) 3.
27. T. Ioannides, G. R. Gavalas, *J. Membr. Sci.* **77** (1993) 207.
28. D. Casanave, A. G. Fendler, J. Sanchez, R. Loutaty, J. A. Dalmon, *Catal. Today* **25** (1995) 309.
29. D. Casanave, P. Ciavarella, K. Fiaty, J. A. Dalmon, *Chem. Eng. Sci.* **54** (1999) 2807.
30. Y. Zhu, R. G. Minet, T. T. Tsotsis, *Cat. Lett.* **18** (1993) 49.
31. G. R. Gallaher, Jr., T. E. Gerdes, P. T. K. Liu, *Sepn. Sci. and Technol.* **28** (1993) 309.
32. F. Tiscareno-Lechuga, C. G. Hill, Jr., M. A. Anderson, *Appl. Catal. A: General* **96** (1993) 33.
33. F. Tiscareno-Lechuga, C. G. Hill, Jr., M. A. Anderson, *J. Membr. Sci.* **118** (1996) 65.
34. V. M. Gryaznov, V. S. Smirnov, *Kinet. Catal.* **18** (1977) 485.
35. US 4 388 479, 1983(V. M. Gryaznov, A. N. Karavanov, T. M. Belosljudova, A. M. Ermolaev, A. P. Maganjunk, I. K. Sarycheva).
36. V. M. Gryaznov, A. N. Karavanov, *Khim. Farm. Zh.* **13** (1979) 74.
37. Itoh, N, K. Haraya, *Catal. Today* **56** (2000) 103.
38. N. Itoh, W. C. Xu, S. Hara, K. Sakaki, *Catal. Today* **56** (2000) 307.
39. K. Otsuka, T. Yagi, *J. Catal.* **145** (1994) 289.
40. E. Panagos, L. Voudouris, M. Stoukides, *Chem. Eng. Sci.* **51** (1996) 3175.

41. M. M. Ermilova, N. V. Orekhova, V. M. Gryaznov: "Optimization of the Selective Hydrogenation Process by Membrane Catalysts", in R. Bredesen (ed.):. *Proc. Fourth Workshop: Optimisation of Catalytic Membrane Reactors Systems*, Oslo, Norway, May, 1997, pp. 187 – 189.
42. P. Cini, M. P. Harold, *AIChE J.* **37**(1991) no. 7, 997.
43. M. Torres, J. Sanchez, J. A. Dalmon, B. Bernauer, J. Lieto, *Ind. Eng. Chem. Res.* **33** (1994) 2421.
44. O. Monticelli, A. Bezzi, A. Bottino, G. Capannelli, A. Servida in "Hydrogenation of Cinnamaldehyde: the Use of Three Phase Catalytic Membrane Reactors", *Fourth Workshop: Optimisation of Catalytic Membrane Reactors Systems*, Oslo, Norway, May, 1997, p. 109.
45. B. C. H. Steele, *Mater. Sci. Eng.* **B13** (1992) 79.
46. M. Stoukides, *Ind. Eng. Chem. Res.* **27** (1988) 1745.
47. Y. Teraoka, H. M. Zhang, S. Furukawa, N. Yamazoe, *Mater. Res. Bull.* **33** (1988) 51.
48. M. Schwartz, J. H. White, M. G. Myers, S. Deych, A. F. Sammells, in "The Use of Ceramic Membrane Reactors for the Partial Oxidation of Methane to Synthesis Gas", *Preprints 213th ACS National Meeting*, San Francisco, CA, April 13 – 17, 42, (1997).
49. A. F. Sammells, M. Schwartz, R. A. Mackay, T. F. Barton, D. R. Peterson, *Catal. Today* **56** (2000) 325.
50. W. Wang, Y. S. Lin, *J. Membr. Sci.* **103** (1995) 219.
51. E. E. Wolf, *Methane Conversion by Oxidative Processes,* Reinhold, New York 1992.
52. H. J. M. Bouwmeester, J. E. ten Elshof in "Oxidative Coupling of Methane in a Dense Ceramic Membrane Reactor", *Fourth Workshop: Optimization of Catalytic Membrane Reactors Systems*, Oslo, Norway, May, 1997, p. 27.
53. Y. Zeng, Y. S. Lin in ″Experimental Study of Oxidative Coupling of Methane in Dense Ceramic Membrane Reactor,″ *Proc. 5th ICIM*, Nagoya, June 22 – 26, 1998, p. 354.
54. I. P. Androulakis, S. C. Reyes, *AIChE J.* **45** (1999) 860.
55. D. Lafarga, J. Santamaria, M. Menendez, *Chem. Eng. Sci.* **49** (1994) 2005.
56. J. Coronas, M. Menendez, J. Santamaria, *Chem. Eng. Sci.* **49** (1994) 2015a.
57. J. Coronas, M. Menendez, J. Santamaria, *Chem. Eng. Sci.* **49** (1994) 4749b.
58. J. Herguido, D. Lafarga, M. Menendez, J. Santamaria, C. Guimon, *Catal. Today* **25** (1995) 263.
59. Y. H. Ma, Y. Lu, A. G. Dixon, W. R. Moser in "A Study of the Oxidative Coupling of Methane in a Lanthanum Stabilized Porous γ-Al_2O_3 Membrane Reactor", *Proc. 5th ICIM*, Nagoya, June 22 – 26, p. 330, 1998.
60. D. Eng, M. Stoukides in "Partial Oxidation of Methane in a Solid Electrolyte Cell,″ *Proc. 9th Int. Congr. Catalysis Chem. Inst. of Canada*, Ottawa, 2, 1988, 974.
61. D. Eng, M. Stoukides, *Catal. Rev. Sci. Eng.* **33** (1991) 375.
62. J. T. Mazanec, T. L. Cable, J. G. Frye, Jr. *Solid State Ionics* **53** (1992) 111.
63. EP-A 0 399 833, 1990 (T. L. Cable, T. J. Mazanec, J. G. Frye).
64. J. T. Mazanec in H. U. Anderson, A. C. Khandkar, M. Liu (eds): "Electropox Gas Reforming", *Proc. First Int. Symp. on Ceramic Membranes* 16 (1997) 95 – 24.
65. U. Balachandran, J. T. Dusek, S. M. Sweeney, R. B. Poeppel, R. L. Mieville, P. S. Maiya, M. S. Kleefisch, S. Pei, T. P. Kobylinsky, C. A. Udovich, *Ceram. Bull.* **74** (1995) 71.
66. U. Balachandran, J. T. Dusek, R. L. Mieville, R. B. Poeppel, M. S. Kleefisch, S. Pei, T. P. Kobylinsky, C. A. Udovich, A. C. Bose, *Appl. Catal. A* **133** (1995) 19.
67. U. Balachandran, J. T. Dusek, P. S. Maiya, B. Ma, R. L. Mieville, M. S. Kleefisch, C. A. Udovich, *Catal. Today* **36** (1997) 265.
68. U. Balachandran, P. S. Maiya, B. Ma, J. T. Dusek, R. L. Mieville, J. J Picciolo: "Development of Mixed-Conducting Ceramic Membranes for Converting Methane to Syngas", *Proceedings of the Fifth International Conference on Inorganic Membranes*, Nagoya, June 22 – 26, 1998.
69. P. S. Maiya, T. J. Anderson, R. L. Mieville, J. T. Dusek, J. J. Picciolo, U. Balachandran, *Appl. Catal. A-Gen.* **196** (2000) 65.
70. T. J. Mazanec, personal communication, 1999.
71. P. V. Hendriksen, P. H. Larsen, M. Mogensen, F. W. Poulsen, K. Wiik, *Catal. Today* **56** (2000) 283.
72. J. Yun, H. Hahm, M. Alibrando, E. E. Wolf in "Methane Partial Oxidation in a Fast Flow Non-Selective Membrane Reactor", *III International Conference on Catalysis in Membrane Reactors*, Copenhagen, paper O4, Sept. 8 – 10, 1998.

73. G. Langhendries, G. Baron, I. F. J. Vankelecom, R. F. Parton, P. Jacobs, *Catal. Today* **56** (2000) 131.
74. I. F. J. Vankelecom, P. A. Jacobs, *Catal. Today* **56** (2000) 147.
75. S. Q. Wu, J. E. Gallot, M. Bousmina, C. Bouchard, S. Kaliaguine, *Catal. Today* **56** (2000) 113
76. K. Otsuka, S. Yokoyama, A. Morikawa, *Chem. Lett.* **1985**, 319.
77. US 4 791 079, 1988 (E. A. Hazburn).
78. H. Nagamoto, K. Hayashi, H. Inoue, *J. Catal.* **126** (1990) 671.
79. T. Nozaki, O. Yamazaki, K. Omata, K. Fujimoto, *Chem. Eng. Sci.* **47** (1992) 2945.
80. M. P. Harold, C. Lee, A. J. Burggraaf, K. Keizer, V. T. Zaspalis, R. S. A. de Lange, *Mat. Res. Soc. Bull.* (1994) no. 4, 34.
81. T. Tsunoda, T. Hayakawa, Y. Imai, T. Kameyama, K. Takemira, K. Fukuda, *Catal. Today* **25** (1995) 371.
82. T. Hibino, T. Sato, K. I. Ushiki, Y Kuwahara, *J. Chem. Soc. Faraday Trans.* **91** (1995) 4419.
83. J. Langguth, R. Dittmeyer, H. Hofmann, G. Tomandl, *Appl. Catal. A: General* **158** (1997) 287.
84. Y. Zeng, Y. S. Lin, *Ind. Eng. Chem. Res.* **36** (1997) 277a.
85. Y. Zeng, Y. S. Lin, *Appl. Catal. A: General* **159** (1997) 101b.
86. O. Garnier, J. Shu, B. P. A. Grandjean in D. E. Fain (ed.): "Catalytic Conversion of Methane into Hydrogen Higher Hydrocarbons in a Pd-based Membrane Reactor", *Proc. 4th ICIM, Gatlinburg, TN*, July 14 – 18, 1998, p. 334.
87. H. Borges, A. G. Fendler, C. Mirodatos, P. Chanaud, A. Julbe, *Catal. Today* **25** (1995) 377.
88. Y. P. Lu, A. G. Dixon, W. R. Moser, Y. H. Ma, U. Balachandran, *Catal. Today* **56** (2000) 297.
89. S. Hamakawa, T. Hayakawa, K. Suzuki, R. Shiozaki, K. Takehira, *Denki Kagaku* **65** (1997) 1049b.
90. H. Dong, G. X. Xiong, Z. P. Shao, S. L. Liu, W. S. Yang, *Chin. Sci. Bull.* **45** (2000) 224.
91. H. Zhong-Tao, L. Ru-Xuan, paper presented at the 1st Int. Workshop on Catalytic Membranes, Sept. 26 – 28,1994, Lyon-Villeurbanne, France.
92. C. Yang, N. Xu, J. Shi, in "Numerical Simulation Experimental Study of Catalytic Inert Membrane Reactor for Partial Oxidation of CH_4 to HCHO", *Proc. 5th ICIM*, Nagoya, Japan, 1998, p. 616.
93. G. Lu, S. Shen, R. Wang, *Catal. Today* **30** (1996) 41.
94. Q. Liu, J. Rogut, B. Chen, J. L. Falconer, R. D. Noble, *Fuel* **75** (1996) 1748.
95. A. L. Y. Tonkovich, R. Secker, E. Reed, E. Roberts, J. Cox, *Sepn. Sci. Technol.* **30** (1995) 397.
96. A. L. Y. Tonkovich, J. L. Zilka, D. M. Jimenez, G. L. Roberts, J. L. Cox, *Chem. Eng. Sci.* **51** (1996) 89.
97. S. Hamakawa, T. Hayakawa, K. Suzuki, M. Shimizu, K. Takehira in R. K. Grasseli, S. T. Oyama, A. M. Gaffney, J. E. Lyons (eds.): "Electrochemical Oxidation of Propene Using a Membrane Reactor with Solid Electrolyte", *Proc. 3rd Word Congress on Oxidation Catalysis*, 1997, 1323.
98. A. P. E. York, S. Hamakawa, T. Hayakawa, K. Sato, T. Tsunoda, K. Takehira, *J. Chem. Soc. Faraday Trans.* **92** (1996) 3579.
99. M. Pena, D. M. Carr, K. L. Yeung, A. Varma, *Chem. Eng. Sci.* **53** (1998) 3821.
100. D. Lafarga, A. Varma, *Chem. Eng. Sci.* **55** (2000) 749.
101. A. Pantazidis, J. A. Dalmon, C. Mirodatos, *Catal. Today* **25** (1995) 403.
102. G. Capannelli, E. Carosini, F. Cavani, O Monticelli, F. Trifiro, *Chem. Eng. Sci.* **51** (1996) 1817.
103. M. J. Alfonso, A. Julbe, D. Farrusseng, M. Menendez, J. Santamaria, *Chem. Eng. Sci.* **54** (1999) 1265.
104. A. Julbe, D. Farrusseng, J. C. Jalibert, C. Mirodatos, C. Guizard, *Catal. Today* **56** (2000) 199.
105. R. Ramos, M. Menendez, J. Santamaria, *Catal. Today* **56** (2000) 239.
106. C. Tellez, M. Menendez, J. Santamaria, *AIChE J.* **43** (1997) 777.
107. T. M. Raybold, M. C. Huff, *Catal. Today* **56** (2000) 35.
108. R. Mallada, M. Menendez, J. Santamaria, *Catal. Today* **56** (2000) 191a.
109. R. Mallada, M. Pedernera, M. Menendez, J. Santamaria, *Ind. Eng. Chem. Res.* **39** (2000) 620b.
110. Y. H. Ma, W. R. Moser, S. Pien A. B. Shelekhin, "Experimental Study of H2S Decomposition in a Membrane Reactor", in Y. H. Ma (ed.): *Proc. 3rd ICIM*, July 10 – 14, 1994, Worcester, MA, pp. 281 – 286.
111. D. J. Edlund, W. A. Pledger, *J. Membr. Sci.* **94** (1994) 111.
112. J. P. Collins, J. D. Way, *J. Membr. Sci.* **96** (1994) 259.

113. E. N. Gobina, J. S. Oklany, R. Hughes, *Ind. Eng. Chem. Res.* **34** (1995) 3777.
114. C. C. Chang, C. M. Reo, C. R. F. Lund, *Appl. Catal. B - Envir.* **20** (1999) 309.
115. V. T. Zaspalis, W. van Praag, K. Keizer, J. G. van Ommen, J. R. H. Ross, A. J. Burggraaf, *Appl. Catal.* **74** (1991) 249b.
116. H. J. Sloot, C. A. Smolders, W. P. M. van Swaaij, G. F. Versteeg, *AIChE J.* **37** (1991) 997.
117. H. J. Sloot, G. F. Versteeg, W. P. M. van Swaaij, *Chem. Eng. Sci.* **45** (1990) 2415.
118. J. M. Veldsink, G. F. Versteeg, W. P. M. van Swaaij, paper presented at the 1st Int. Workshop on Catalytic Membranes, Lyon-Villeurbanne, France, Sept. 26 – 28, 1994.
119. H. W. J. P. Neomagus, W. P. M. van Swaaij, G. F. Versteeg, *J. Membr. Sci.* **148** (1998) 147.
120. M. P. Pina, M. Menendez, J. Santamaria, *Appl. Catal. B.* **11** (1996) L19a.
121. M. P. Pina, S. Irusta, M. Menendez, J Santamaria, R. Hughes, N. Boag, *Ind. Eng. Chem. Res.* **36** (1996) 4557b.
122. S. Irusta, M. P. Pina, M. Menedez, J. Santamaria, *Catal. Lett.* **54** (1998) 69.
123. W. A. Jacoby, P. C. Maness, D. M. Blake, E. J. Wolfrum in "Photocatalytic Membrane Reactor for Removal of Chemicals Bioaerosols from Indoor Air", *III International Conference on Catalysis in Membrane Reactors*, Copenhagen, paper O13, Sept. 8 – 10, 1998.
124. E. Kikuchi, *Catal. Today* **56** (2000) 97.
125. US 07 965 011; CA 2 081 170, 1992 (A. M. Adris, J. R. Grace, C. J. Lim, S. S. E. H. Elnashaie).
126. A. M. Adris, C. J. Lim, J. R. Grace, *Chem. Eng. Sci.* **49** (1994) 5833.
127. A. M. Adris, J. R. Grace, *Ind. Eng. Chem. Res.* **36** (1997) 45.
128. K. Hou, M. Fowles, R. Hughes, *Catal. Today* **56** (2000) 13.
129. K. Aasberg-Petersen, C. S. Nielsen, S. L. Jorgensen, *Catal. Today* **46** (1998) 193.
130. Hojlund P. E. Nielsen, K. Aasberg-Petersen, S. Laegsgaard Jorgensen in "A Review on Hydrogen Selective Membranes – Obtained Results Future Perspectives", *III International Conference on Catalysis in Membrane Reactors*, Copenhagen, paper O26, Sept. 8 – 10, 1998.
131. Y. M. Lin, G. L. Lee, M. H. Rei, *Catal. Today* **44** (1998) 343.
132. S. Hara, W. C. Xu, K. Sakaki, N. Itoh, *Ind. Eng. Chem. Res.* **38** (1999) 488.
133. M. Bracht, P. T. Alderliesten, R. Kloster, R. Pruschkek, G. Haupt, E. Xue, J. R. Ross, M. K. Koukou, N. Papayannakos, *Energy Conversion and Management* **38** (1997) 159.
134. V. Violante, A. Basile, E. Drioli, *Fusion Engineering and Design* **22** (1993) 257.
135. J. Fan, H. Ohya, H. Ohashi, M. Aihara, T. Takeuchi, Y. Negishi, S. I. Semenova, *J. Membr. Sci.* **162** (1999) 125.
136. V. T. Zaspalis, W. van Praag, K. Keizer, J. G. van Ommen, J. R. H. Ross, A. J. Burggraaf, *Appl. Catal.* **74** (1991) 235a.
137. G. Saracco, V. Specchia, *Chem. Eng. Sci.* **55** (2000) 897.
138. S. Uemiya, N. Sato, H. Ando, T. Matsuda, E. Kikuchi, *Appl. Catal.* **67** (1991) 223.
139. E. Kikuchi, "Palladium Ceramic Membranes for Selective Hydrogen Permeation Their Application to Membrane Reactor", in *Proc.1st Int. Workshop on Catalytic Membranes*, Sept. 26 – 28, 1994, Lyon-Villeurbanne, France.
140. E. Kikuchi, Y. Nemoto, M. Kajiwara, S. Uemiya, T. Kojima, *Catal. Today* **56** (2000) 75.
141. R. G. Minet, S. P. Vasileiadis, T. T. Tsotsis, in G. A. Huff, D. A. Scarpiello (eds.): "Experimental Studies of a Ceramic Membrane Reactor for the Steam/Methane Reaction at Moderate Temperatures, 400 – 700 °C", *Proc. of Symp. on Natural Gas Upgrading*, **37** (1992) 245.
142. L. Mleczko, T. Ostrowski, T. Wurzel, *Chem. Eng. Sci.* **51** (1996) 3187.
143. G. Barbieri, V. Violante, P. Di Maio, A. Criscuoli, E. Drioli, *Ind. Eng. Chem. Res.* **36** (1997) 3369.
144. T. M. Raybold, M. C. Huff in "CO_2 Reforming of CH_4 Over Supported Noble Metal Catalysts in a Pd Membrane Reactor", in *Proc. 16th Meeting of the North American Catalysis Society*, Boston, MA, paper C-028, May 30 to June 4, 1999.
145. A. K. Prabhu, S. T. Oyama, *Chem. Lett.* **3** (1999) 213.
146. F. Lipnizki, R. W. Field, P. K. Ten, *J. Membr. Sci.* **153** (1999) 183.
147. F. Lipnizki, S. Hausmanns, G. Laufenberg, R. Field, B. Kunz, *Chem. Ing. Tech.* **70** (1998) 1587.
148. H. E. A. Bruschke, G. Erlington, W. H. Schneider, *Proc. 5th Int. Conf. on Pervap. Proc. in the Chem. Ind.*, R. Brakish (ed.), Brakish Material Corporation, Englewood, NJ, USA, 310, 1991.

149. N. Kitta, S. Sasaki, K. Tanaka, K. I. Okamoto, M. Yamamoto, *Chem. Lett.* **1988**, 2025.
150. S. J. Kwon, K. M. Song, W. H. Hong, J. S. Rhee, *Biotechnol. Bioeng.* **46** (1995) 393.
151. R. Waldburger, F. Widmer, W. Heinzelmann, *Chem. Eng. Tech.* **66** (1994) 850.
152. S. Bitterlich, M. Meissner, W. Hefner in R. Barkish (ed.): *Proc. 6th Int. Conf. on Pervap. Proc. in the Chem. Ind.*, Brakish Material Corporation, Englewood, NJ, 372, 1991.
153. H. H. Nijhuis, A Kemperman, J. T. P. Derksen, F. P. Cuperus in R. Brakish (ed.): *Proc. 6th Int. Conf. on Pervap. Proc. in the Chem. Ind.*, Brakish Material Corporation, Englewood, NJ, 368, 1992.
154. N. Qureshi, H. P. Blaschek, *Biotechnol. Prog.* **15** (1999) 594.
155. C. R. Staudt-Bickel, N. Lichtenthaler, *J. Membr. Sci.* **11** (1996) 135.
156. C. R. Staudt-Bickel, N. Lichtenthaler, *Prog. Intl. Cong. Org. Membr.* **96** (1996) 394.
157. W. J. W. Bakker, I. A. A. C. M. Bos, W. L. P. Rutten, J. T. F. Keurentjes, M. Wessling *Proc. 5th Intl. Congress Inorg. Membr.*, Nagoya, Japan, 448, 1998.
158. X. Ni, Z. Xu, Y Shi, Y. Mu, *Water Treat.* **10** (1995) 115.
159. Y. Zhu, R. G. Minet, T. T. Tsotsis, *Chem. Eng. Sci.* **51** (1996) 4103.
160. M. Mulder: *Basic Principles of Membrane Technology,* Kluwar, Dortrecht 1996.
161. A. S. Michaels, *Chem. Eng. Prog.* **64** (1968) 31.
162. US 4 795 704, 1989 (S. L. Matson).
163. H. Moueddeb, J. Sanchez, C. Bardot, M. Fick, *J. Membr. Sci.* **114** (1996) 59.
164. J. P. S. G. Crespo, A. M. R. B. Xavier, M. T. O. Barreta, L. M. D. Goncalves, J. S. Almeida, M. J. T. Carrondo, *Chem. Eng. Sci.* **47** (1992) 205.
165. M. Cheyran, M. A. Mehaia, *Proc. Biochem.* **19** (1984) 204.
166. US 4 945 047, 1990 (R. Legoux, P. Lepatois, J. E. Liauzun, B. Niaudet, W. G. Roskam, W. Roskam).
167. B. Bauer, H. Chmel, F. Effenberger, H. Starthmann, *Proc. Int. Cong. on Membr. and Membr. Processes, ICOM*, Heidelberg, Germany, August 1993.
168. A. M. R. B. Xavier, L. M. D. Gonçalves, J. L. Moreira, M. J. T. Carrondo, *Biotechnol. Bioeng.* **45** (1995) 320.
169. Y. Koyama, K. Shimazaki, K. Akashi, Y. Kawahara, K. Kubota, H. Yoshii, in O. M. Neijssel, R. R. van der Meer, K. Ch. A. M Luyben, Eds, "Production of L-Aspartic Acid by Membrane Reactor", in *Proc. 4th Eur. Congr. in Biotechn.*, Elsevier, Amsterdam, 1 (1987) 119.
170. A. Bosetti, D. Bianchi, N. Andriollo, D. Cidaria, P. Cesti, G. Sello, P. Di Gennaro, *Chem. Engng. J.* **66** (1996) 375.
171. M. K. H. Liew, A. G. Fane, P. Rogers, *J. Chem. Tech. Biotechnol.* **64** (1996) 200.
172. S. D. Doig, A. T. Boam, D. I. Leak, A. G. Livingston, D. C. Stuckey, *Biotechnol. Bioengng.* **58**(1998) no. 6, 587.
173. Y. Shi, J. Ploof, A. Correia, *IVD Technology Mag.* 1999 no. 5.
174. J. Hopkinson, *Bio/technol.* **3** (1985) 225.
175. R. Kurkela, E. Fraune, P. Vihko, *Biotechniques* **15** (1993) 674.
176. J. M. Piret, C. L. Cooney, *Biotechnol. Bioengng.* **36** (1990) 902.
177. J. Feder, R. Tolber, *Int. Biotechnol. Lab.* **1985**, no. 6, 41.
178. F. Molinari, F. Aragozzini, J. M. S. Cabral, D. M. F. Prazeres, *Enzym. Microb. Technol.* **20** (1997) 604.
179. FP 2 597 114, 1987 (R. Legoux, P. Lepatois, J. E. Liauzun, B. Niaudet, W. G. Roskam, W. Roskam).
180. A. Nipkow, J. G. Zeikus, P. Gerhardt, *Biotechnol. Bioeng.* **34** (1989) 1075.
181. T. L. Evans, R. A. Miller, *BioTechniques* **1988**, 6.
182. B. Y. Chung, H. N. Chang, I. H. Kim, *Enzyme Microbiol. Technol.* **9** (1987) 345.
183. D. M. F. Prazeres, J. M. S. Cabral, *Enzyme Microb. Technol.* **16** (1994) 738, 1994.
184. J. M. Engasser, *Biotechnologie, Technique et Documentation Lavoisier* , Paris 1988.
185. C. Wandrey, R. R. van der Meer, K. Ch. A. M. Luyben (eds.): *Proc. 4th European Congress in Biotechnology* Elsevier, Amsterdam 1987.
186. E. Schimidt, W. Hummel, C. Wandrey in O. M. Neijssel, R. R van der Meer, K. Ch. A. M. Luyben (eds.): "Continuous production of L-Phenylalanine from Acetomidocinnamic Acid using the Three-Enzyme-System Acylase/L-PHEDH/FDH", *Proc. 4th Eur. Congr. in Biotechn.*, Elsevier, Amsterdam, 2 (1987) 189.
187. M. Biselli, U. Kragl, C. Wandrey in K. Drauz, H. Waldmann (eds.): *Enzyme Catalysis in Organic Synthesis* VCH, Weinheim 1995.
188. U. Kragl, D. Vasic-Racki, C. Wandrey, *Bioprocess Engng.* **14** (1996) 291.

189. J. Bongs, D. Hahn, U. Schorken, G. A. Sprenger, U. Kragl, C. Wandrey, *Biotechnol. Lett.* **19** (1997) 213.
190. A. Liese, T. Zelinski, M. R. Kula, H. Kierkels, M. Karutz, U. Kragl, C. Wandrey, *J. Mol. Catal. B: Enzymatic* **4** (1998) 91.
191. M. M. Hoq, T. Yamane, S. Shimizu, T. Funada, S. Ishida, *J. Am. Oil Chem. Soc.* **62** (1985) no. 6, 1016.
192. L. Giorno, R. Molinari, M. Natoli, E. Drioli, *J. Membr. Sci.* **125** (1997) 177.
193. Q. Gan, F. Baykara, H. Rahmat, L. R. Weatherley, *Catal. Today* **56** (2000) 179.
194. H. G. Hicke, M. Ulbricht, M. Becker, S. Radosta, A. Heyer, *J. Membr. Sci.* **161** (1999) 239.
195. J. L. Lopez, S. L. Matson, *J. Membr. Sci.* **125** (1997) 189.
196. M. Furui, T. Furutani, T. Shibatani, Y. Nakamoto, T. Mori, *J. Ferm. Bioengng.* **81** (1996) no. 1, 21.
197. R. Wichmann, C. Wandrey, A. F. Buckmann, R. Kula, *Biotechnol. Bioengng.* **23** (1981) 2789.
198. W. Leuchtenberger, M. Karrenbauer, U. Plocker, *Ann. N.Y. Acad. Sci.* **434** (1984) 78.
199. C. Fonade, M. Y. Jaffrin in G. Dauphin, F. Rene, P. Aimar (eds.): "Regimes Instationnaires: Mise en Cuvre des Perturbations", Les Separations par membrane dans les Procedes de l'Industrie Alimentaire, Technique et Documentation Lavoisier Paris, Chap. 4, 1998.
200. J. S. Devinny, M. A. Deshusses, T. S. Webster: *Biofiltration for Air Pollution Control,* Lewis/CRC Press, Boca Raton FI, 1998.
201. M. Gander, B. Jefferson, S. Judd, *Sep. Purif. Technol.* **18** (2000) 119.
202. M. Praderie, H. Buisson, H. Paillard T. Vouillon, "Le Traitement des Eaux Residuaires par le Procede BIOSEP, Boues a Membrane Immergées", *Galvano-organo-traitements de surface* **685** (1998) 390.
203. S. Churcose, "Operating Experiences with the Kubota Submerged Membrane Activated Sludge Process", *Proc. 1st Int. Meeting on Membrane Bioreactors for Wastewater Treatment*, Cranfield University, 5 – 6 March, 1997.

Polyureas

Constantin I. Chiriac, Institute of Macromolecular Chemistry "Petru Poni", Jassy, Romania
Fulga Tanasă, Institute of Macromolecular Chemistry "Petru Poni", Jassy, Romania

1. Introduction 1
2. Synthesis 1
3. Polyureas as Starting Materials for Other Polymers 6
4. Homopolyureas 7
5. Copolyureas 8
5.1. Simple Copolyureas 8
5.2. Copolyureas Containing Other Functional Groups 9
6. Structure 11
7. Properties 13
7.1. Physical Properties 13
7.2. Chemical Properties 13
8. Uses 15
8.1. Foams 15
8.2. Moldings 17
8.3. Fibers 18
8.4. Films and Membranes 18
8.5. Coatings 19
8.6. Biomedical Applications 19
8.7. Other Applications 20
9. Safety and Environmental Aspects . . 20
10. References 20

1. Introduction

Polyureas are defined as polymers which contain ureylene groups –NHCONH– in the polymer chain. Urea, regarded as the diamide of carbonic acid, was first synthesized by Wöhler in 1892, but the chemistry and technology of polyureas have a more recent origin.

Linear polyureas are thermoplastic polycondensation products with aromatic (R = arylene) or aliphatic (R = alkylene) structures:

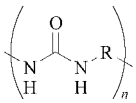

Polyureas or copolyureas with aliphatic structures exhibit a difference of 50 – 100 °C between the softening point and the onset of decomposition. They can be used as moldings. Polyureas and copolyureas with aromatic structures have softening points near their decomposition temperatures and are soluble in organic solvents, particularly in aprotic dipolar solvents such as DMF, *N*-methylpyrrolidinone (NMP), and DMSO. They are used for the preparation of lacquers, varnishes, and coatings [1]. The relationship between the softening points and structure is shown in Table 1.

Polyureas were first produced on a commercial scale at I.G. Farben by employing the reaction between diisocyanates and diamines. Mitsui Toatsu Chemical produces synthetic fibers from nonamethylenediamine and urea. Polyureas and copolyureas, particularly poly(urethaneurea)s, have many practical applications as foams, elastomers, fibers, etc.

Worldwide consumption was below 4.8×10^6 t in 1984 (in Europe, 2×10^6 t) and 5.9×10^6 t in 1995 [4]. More than 75 % of consumption was for foams [5]. Growth is estimated at 3 – 5 % per annum in the period 1995 to 2000. The number of products and producers continues to increase, and new processing technologies for thermosets, such as reaction injection molding (RIM), liquid injection molding (LIM), and others, are being developed.

2. Synthesis

The synthesis of polyureas is based on reactions between polyamines, especially diamines,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.d21_d01

Table 1. Alternating copolyureas derived from diamines and diisocyanates

Diamine	Diisocyanate	Intrinsic viscosity, dL/g	Softening point, °C	Ref.
Piperazine	1,4-cyclohexylene diisocyanate	0.60	324	[2]
2,5-Dimethylpiperazine	1,4-cyclohexylene diisocyanate	0.48	390	[2]
2,5-Dimethylpiperazine	1,2-bis(4-isocyanatophenyl)ethene	0.33	362 (decomp.)	[2]
2,5-Dimethylpiperazine	bis(4-isocyanatophenyl)methane	1.30	350 (decomp.)	[2]
1,4-Phenylenediamine	bis(4-isocyanatophenyl)methane	0.36	400 (decomp.)	[2]
Hexamethylenediamine	4-methyl-1,3-phenylene diisocyanate	0.47	225 – 230	[3]

with aliphatic, aromatic, or heterocyclic structures and carbon dioxide, carbonyl sulfide, carbonic esters, phosgene, urea, urethane, and isocyanates.

From Diisocyanates and Diamines. Polyureas are readily prepared by a polyaddition reaction between diisocyanates and diamines.

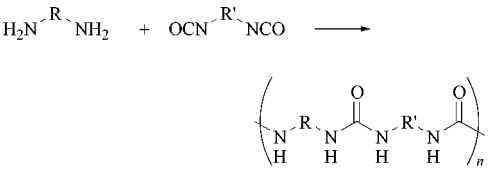

Isocyanates are reactive products, and the reactivity toward compounds containing active hydrogen is as follows: $RNH_2 > ArNH_2 >$ primary OH > water > secondary OH > tertiary OH > phenolic OH > COOH > RNHCONHR′ > RCONHR′ > RNHCOOR′ [6]. An isocyanate group bonded to an aromatic ring is more reactive than one bonded to an aliphatic chain. Many aliphatic amines are extremely reactive with isocyanates, even at low temperatures. Relative reactivities of amines with phenyl isocyanate in diethyl ether at 0°C [7] follow:

NH_3	1.00
Ethylamine	9.72
n-Propylamine	8.22
n-Butylamine	9.17
n-Amylamine	9.17
Aniline	0.53

Secondary aliphatic and primary aromatic amines react similarly. Secondary aromatic amines are less reactive [7].

For the amine – isocyanates reaction, the following mechanism has been proposed [8]:

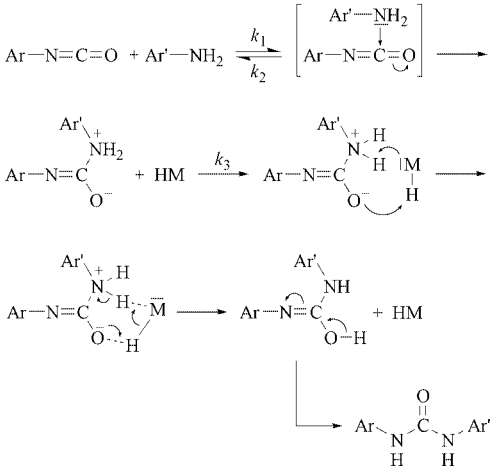

In this mechanism, HM is a proton acceptor/donor, for example, an amine, urea, or carboxylic acid, which acts as catalyst [9]. The reactions between isocyanates and amines are usually autocatalytic, but they can be catalyzed by many other catalysts, amines, or acids [10], [11].

Weak aromatic carboxylic acids show a high catalytic activity [12]. In benzoic acid derivatives, electron-acceptor substituents such as NO_2, Cl, or OH reduce catalytic activity, whereas electron-donor substituents such as alkyl groups increase it; 2-methylbenzoic acid is the most active catalyst.

Lewis bases such as 1,4-diazabicyclo-[2.2.2.]-octane (DABCO), triethylamine, dimethylbenzylamine, and pyridine also exhibit catalytic activity. 4-(Dimethylamino)pyridine is an exceptional catalyst, not only in reactions with amines (strongly nucleophilic reagents), but even with alcohols (weakly nucleophilic reagents) [13]. For reactions with alcohols, the following mechanism has been proposed [14]:

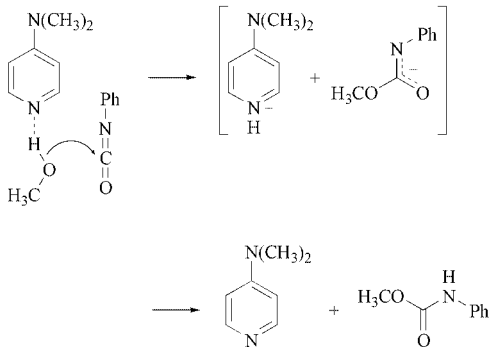

Insertion catalysts such as dibutyltin dilaurate, tin(II) octanoate, and cobalt naphthenate have also been used in this field [15].

The effect of substituents on the reaction between aromatic isocyanates and aromatic amines has been studied [10]. The reaction rates were determined by refluxing 0.005 mol substituted phenyl isocyanate and 0.005 mol substituted aniline for 1 h in a mixture of toluene and benzene. The urea yields are shown in Table 2.

The effect of substituents on the reactivity of the isocyanate decreases in the order: 2,4-$(NO_2)_2$ > 3,5-$(NO_2)_2$ > 4-NO_2 > 3-NO_2 > 2-NO_2 ≫ H, 4-CH_3, 3-OCH_3, 4-OCH_3. Electron-acceptor substituents in aromatic isocyanates increase the reaction rate. The same substituents in aromatic amines reduce the reaction rate. Steric hindrance is also influential. *Ortho* substitution in aromatic isocyanates or amines reduces the reaction rate.

The reaction between diamines and diisocyanates is usually carried out by interfacial polyaddition [16] or in solution [17]. In the bulk, infusible, insoluble cross-linked polymers can be obtained instead of the desired linear polyureas. Cross-linking occurs by the interaction of the highly reactive free isocyanate groups with active hydrogen atoms from the urea groups [18]. The most important application of this reaction is the synthesis of poly(urethaneurea)s, also called copolyureas.

Another interesting application for polyureas is in nonlinear optics (NLO); the NLO response of polyureas is extremely rapid, because effects occur primarily through electronic polarization [19]. It is now well-known that molecules containing electron-donor and electron-acceptor groups separated by a large conjugated π framework possess large values of second-order molecular hyperpolarizability [19]. For the synthesis of polyureas with these properties, 4-(dialkylamino)-4′-(alkylsulfonyl)azobenzene NLO chromophores have been used, which contain methylsulfonyl groups as electron-acceptor groups, 4-dialkylamino groups as electron-donor groups, and between them azo groups [20–22].

Another interesting application is the synthesis of hyperbranched polyureas (Fig. 1) [23], which have unusual properties [24].

In these syntheses, the most important monomer is 3,5-diaminobenzoyl azide (**4**), which on heating in NMP at 110 °C under argon for 16 h gives unstable 3,5-diaminophenylisocyanate (**5**) and finaly hyperbranched polyureas. Curiously, these hyperbranched polyureas are soluble in aprotic dipolar solvents such as DMSO, NMP, and DMF.

From Diisocyanates and Water. Polyureas can be prepared from diisocyanates and water:

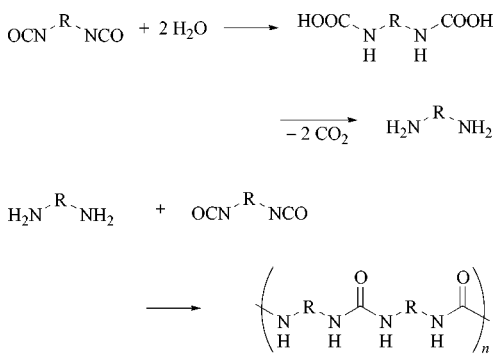

The diamine resulting from reaction of water with the diisocyanate reacts with unconverted diisocyanate more rapidly than water, and polyureas are obtained in high yields. The reactivity of water with diisocyanates is similar to that of secondary alcohols [10]. This apparently simple reaction is, in fact, quite complicated:

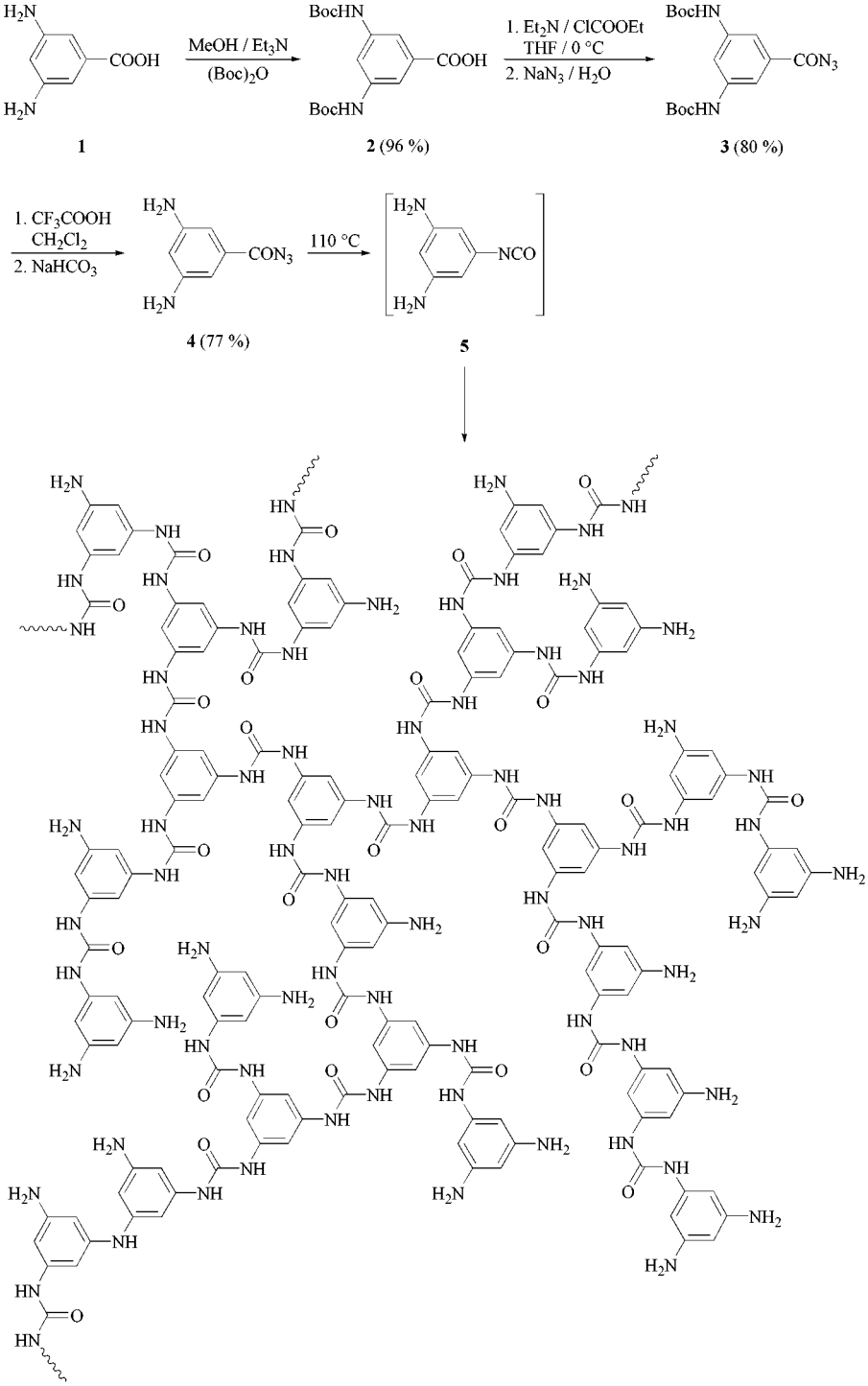

Scheme 1. Synthesis of hyperbranched polyureas

Table 2. Urea yields from aromatic isocyanates and aromatic amines

Isocyanate substituent	Amine substituent								
	2,4-$(NO_2)_2$	2-NO_2	3,5-$(NO_2)_2$	4-NO_2	3-NO_2	3-OCH_3	H	4-CH_3	4-OCH_3
4-OCH_3	–	–	–	–	15	91	90	95	98
4-CH_3	–	–	–	3	14	90	95	98	–
H	0	0	0	0	26	92	92	100	–
3-OCH_3	–	–	–	0	46	95	–	–	–
3-NO_2	0	0.5	4	13	52	96	100	–	–
4-NO_2	0	4	8	14	61	97	100	–	–
3,5-$(NO_2)_2$	0	8	16	54	76	95	100	–	–
2-NO_2	0	0	1	8	60	98	100	–	–
2,4-$(NO_2)_2$	0	45	51	76	84	87	100	–	–

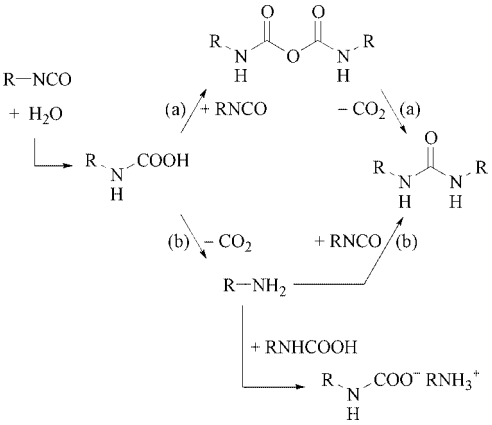

Route (a) is favored when the carbamic acid is stable and reactive toward the isocyanate [25]. Route (b) is favored when the carbamic acid is unstable and decomposes to an amine, which reacts quickly with the isocyanate. Route (c) is favored when the reaction between the carbamic acid and amine is faster than the reaction between the amine and isocyanate. The effect of substituents on the reaction of aromatic isocyanates with water is similar to that observed in the reaction of aromatic isocyanates and amines.

This simple reaction is employed in the manufacture of foams, in which the resulting carbon dioxide acts as a foaming agent. Different amines or amine complexes of metal salts are suitable catalysts for this reaction [26]. With tertiary amines, the following mechanism is proposed [27]:

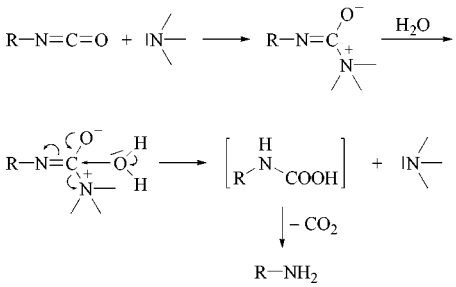

From Diamines and Carbon Dioxide or Its Derivatives. The polycondensation of diamines and carbon dioxide or its derivatives gives polyureas in high yields (X = halogen, OR, NH_2, etc):

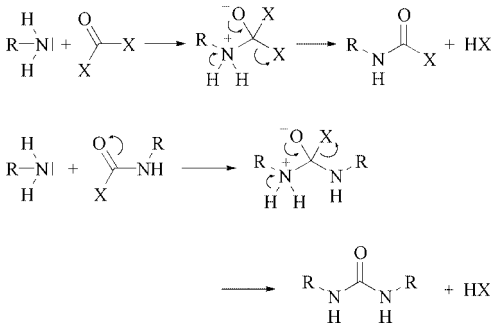

Polyureas are obtained by the polycondensation reaction of *carbon dioxide* and diamines in the presence of diphenyl phosphite in pyridine at 40 °C and 2 MPa. This reaction involves a carbamoyloxy-*N*-phosphonium salt of pyridine as intermediate [28]:

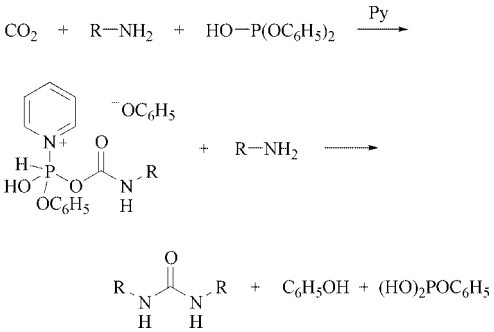

Ethylene chlorophosphites with aliphatic structures are also suitable reagents for this reaction under mild conditions, the mechanism of which involves a four-center transition state [29]:

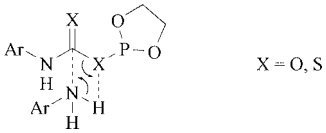

Without catalysts, the reaction of carbon dioxide with diamines requires high temperatures and high pressures [30].

Polycondensation of *phosgene* (carbonic acid dichloride) with diamines gives polyureas [31]. Similarly, *bis(carbamyl chloride)s* prepared from diamines and phosgene, can react with diamines of similar or different structures to afford polyureas (a) or alternating copolyureas (b), respectively:

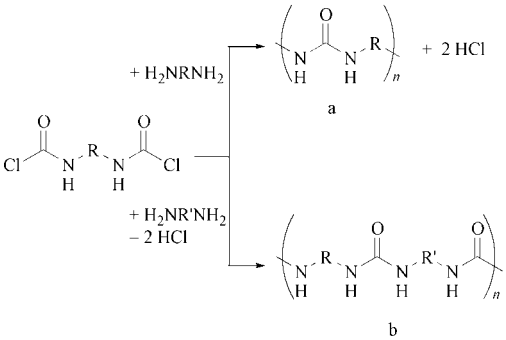

From aliphatic or aromatic diamines and *urea* at high temperatures, polyureas result with the evolution of ammonia. This reaction can be carried out in the melt [32] or in solution [33].

Diphenyl carbonate does not react with aromatic amines in boiling toluene, but the addition of $MgCl_2$ as catalyst in pyridine makes this reaction possible at 120 °C [34]. The same reaction is also possible in phenol or cresol, but at 210 – 230 °C [35]. Activated *carbonic esters*, such as bis(*p*-nitrophenyl)carbonate or bis(2,4-dinitrophenyl)carbonate, can react with diamines at low temperatures to afford polyureas with high viscosities [36]. The rate of reaction is higher for aliphatic diamines than for aromatic diamines [37].

By Other Methods. Diamines can react with *N*, *N′*-carbonyldiimidazole in THF under reflux to give polyureas [38]. Polyureas with moderate molecular masses were prepared by the condensation reaction of lithium carbonate with aromatic diamines in presence of triphenyl phosphite and hexachloroethane in pyridine under reflux [39]. By using the same reaction, polyureas with high thermostability result from 2,8-diaminobenzothiophene or 2,8-diaminophenoxanthine and lithium carbonate [40].

A simple and ingenious method involves a controlled decomposition reaction of sebacoyl diazide in water to form poly(*n*-octylurea) [41].

In an interesting reaction, palladium diacetate catalyzes oxidative polymerization reactions of aliphatic diamines $H_2N(CH_2)_nNH_2$ ($n = 4, 6, 8, 10, 12$) with carbon monoxide in presence of potassium carbonate, triphenyl phosphite, and iodine to afford polyureas [42].

Poly(*N,N′*-diacylurea)s have been obtained by polyaddition reactions of diamides to bis(*N*-acylisocyanate)s [43].

3. Polyureas as Starting Materials for Other Polymers

Polytriketoimidazolidines. Polymers with this structure in the macromolecular chains were prepared by reactions between various polyureas and oxalyl chloride in the presence of pyridine [44]; they decompose 70 – 100 °C higher than the corresponding polyureas.

Polyhydantoins. Aromatic polyhydantoins of high molecular mass have been prepared by a polyaddition/cyclocondensation reaction of diethyl *m*-phenylenebisiminoacetate and aromatic diisocyanates in cresol under reflux [45]:

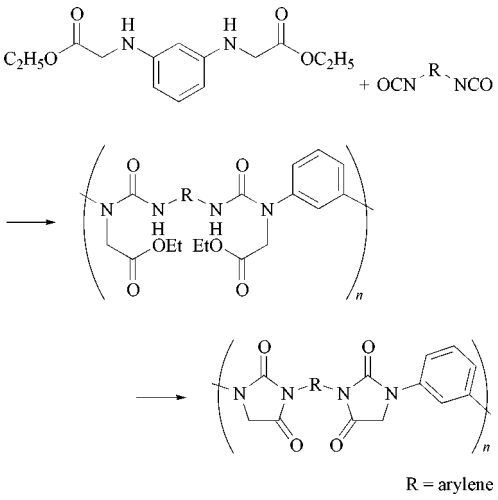

The polyhydantoin obtained from methylenebis(4-phenylisocyanate) is stable in air up to 400 °C. Aspartate polyhydantoin prepolymers can react with blocked polyisocyanates to give products that are useful for coatings, adhesives, and elastomers [46].

Polyquinazolinediones. The reaction of 3,3′-benzidinedicarboxylic acid and diisocyanates with triethylamine as catalyst and NMP as solvent results in polyureas, which are cyclized by heating at 220 °C to polyquinazolinediones [47]:

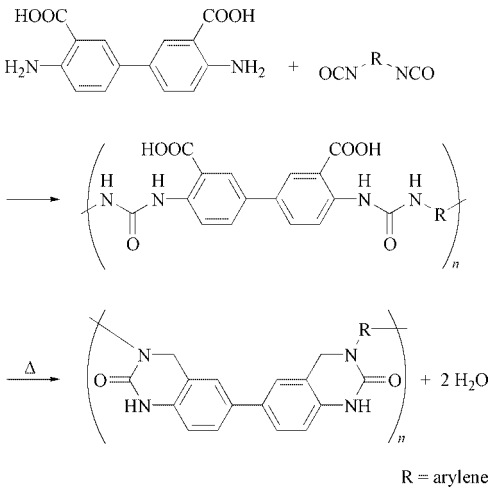

Polyhydrouracils can be obtained by the cyclization of 2-carbomethoxyethyl-substituted polyureas with polyphosphoric acid (PPA) [48]:

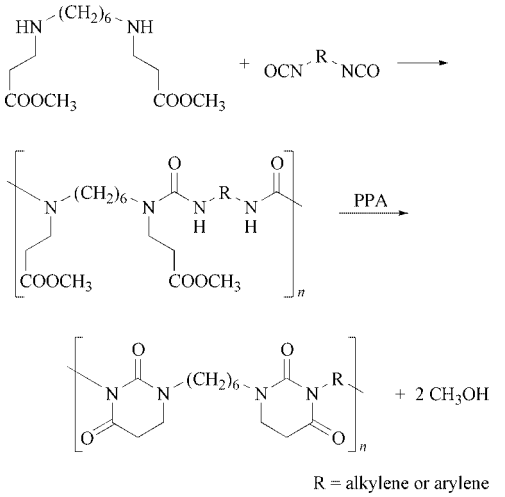

Polyhydrouracils decompose at 400 °C; analogous unsubstituted polyureas decompose at 300 °C.

Other Polymers. Poly(imideurea) elastomers with high elasticity and improved thermostability can be obtained by reaction of polyureas $(R^1NHCONHR^2NHCONH)_n$ with aromatic tetracarboxylic dianhydrides (R^1 = aromatic or aliphatic diisocyanate residue; R^2 = oligomer residue containing amino group-terminated elastomer segment; $n = 2 – 500$) [49].

Heat-resistant polycarbodiimide copolymers for electric cable insulation can be prepared from urea copolymers and 3-methyl-1-phospha-3-cyclopentane-1-oxide as dehydrating agent [50]. The resulting copolymers have high thermostability (at 331 °C, losses are 5 %). Soft Cu conductors were coated with these polycarbodiimides with excellent results.

The end groups of aliphatic polyamides can react with polyureas at high temperatures (200 °C) to give copolymers that exhibit improved fiber dyeability [51].

4. Homopolyureas

Homopolyureas, $[NHCOHN(R)]_n$ have polar structures and high softening points. Homopolyureas with aromatic or heterocyclic structures

have softening points higher than those with aliphatic backbones. The introduction of a symmetrical 1,4-phenylene ring into the polymer chain raises the softening point. *N*-Alkylation lowers softening points. Aliphatic homopolyureas have softening points of about 200 °C and they can be injection-molded, extruded, and processed like other thermoplastics.

5. Copolyureas

Copolyureas have the general formula $[NH(R)NHCONH(R^1)NHCO]_n$, where R and R^1 can be aliphatic, aromatic, or heterocyclic and contain phosphorus, sulfur, silicon, or halogens (simple copolyureas). Copolyureas may contain other functional groups, such as urethane, amide, imide, carbonate, etc.

5.1. Simple Copolyureas

Copolyureas Containing Silicon. Thermostable polysilylureas with good flexibility have been obtained from diamines $H_2N–R^1–NH_2$ and $O[SiR^2(CH_2)_nHCO]_2$ (where $n = 1 – 3$, R^1 = alkyl or aryl, and $R^2 = CH_3$ or C_6H_5) [52]:

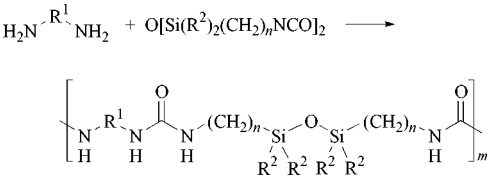

Also, from 1,3-bis(chloromethyldimethyl)siloxane and KOCN the corresponding Si-containing diisocyanate results, which reacts with diamines [53].

Copolyureas Containing Phosphorus. Polyphosphorylureas are prepared from aryl or alkyl phosphoric diisocyanates and diamines by interfacial polyaddition [54]:

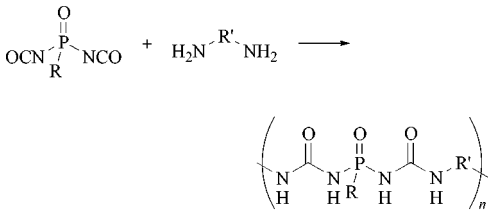

Polyphosphorylureas in which R is C_6H_5 and R′ is *m*-C_6H_4 have a 10 % weight loss at 400 °C and a 40 % weight loss at 600 °C. Flame-retardant copolyureas useful for foams or films are prepared from 1-[(dialkoxyphosphinyl)methyl]-2,4-diaminobenzenes and diisocyanates in NMP at low temperature [55]. By similar methods, phosphorus-containing copolyureas have been prepared from phenyl bis(piperazidophosphate) and diisocyanates [56].

Copolyureas Containing Halogens. Heat-resistant, flame-retardant copolyureas are prepared from fluorinated primary diamines and diisocyanates [57]:

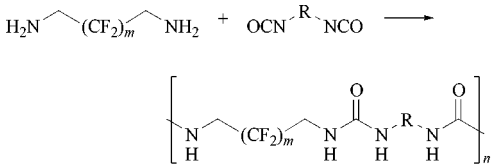

R = alkylene or arylene; *m* = 2 - 4

These copolyureas are useful for oil-resistant coatings and as curing agents for epoxides. Other halogen-containing diamines, such as 1,1,1-trichloro-2,2-bis(4′-chloro-3′-aminophenyl)ethane [58] or 1-(tetrafluoroethoxy)-2,4-phenylenediamine [59] can be used in reaction with diisocyanates, to afford flame-retardant copolyureas.

Copolyureas with Acid or Ester Groups. Polyurea electrolytes are prepared from sodium 3,3′-benzidinedicarboxylate and diisocyanates in the presence of sodium laurate as catalyst [60]:

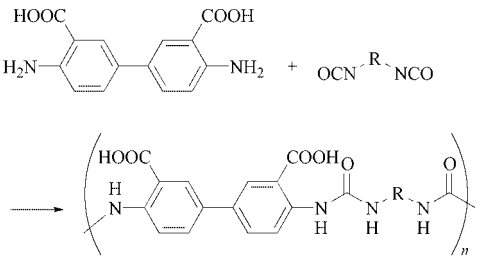

R = alkylene or arylene

From L-lysine diisocyanate methyl ester and optically inactive diamines, optically active copolyureas have been prepared [61]:

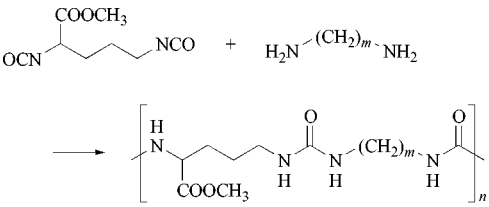

$m = 2 - 6$

Optical studies on these copolyureas revealed that they contain macromolecules with different conformations, ranging from disordered to helical [62]. Treating the polyureas with trifluoroacetic acid causes rupture of hydrogen bonds and reduces the ordering. These polymers can be degraded by chymotripsin or subtilisin [63]. An interesting method for the preparation of these copolyureas involves the polycondensation of L-lysine with 4,4′-dinitrophenylcarbonate [64].

Other Simple Copolyureas. Polycyanoureas with high molecular masses are prepared by a polyaddition of 3,3′-dicyano-4,4′-oxydianiline and diisocyanates in NMP with triethylamine as catalyst [65].

Thermostable copolyureas with oxadiazole rings in the main chain are prepared from bis[5-(3-aminophenyl)-1,3,4-oxadiazol-2-yl]alkanes and diisocyanates in DMF at 60 °C [66].

Polyverdazyl – polyureas are prepared from 1,4-bis[(1,5-diphenyl)-3-(4-aminophenyl)verdazyl-6-yl]butane and diisocyanates [67].

Triazine [68], carborane [69], and ferrocene [70] copolyureas also have high thermostability.

Copolyureas containing crown ethers (cryptates) are used as phase-transfer catalysts for alkylation, esterification, and oxidation. The copolyurea prepared from *trans*-Ar,Ar′-diaminodibenzo-18-crown-6 and TDI has a potassium ion absorption of 45 mg KOH per 100 g polymer, compared to 0.02 mg for the diamine monomer [71].

5.2. Copolyureas Containing Other Functional Groups

Poly(urethaneurea)s are the most important class of copolyureas. They can be prepared from a macrodiol (polyester diols, polyether diols, etc.) and an excess of a diisocyanate; this results in prepolymers with NCO end groups, which react with diamines (chain extenders). At the end of the reaction, a chain terminator (ROH, $RHNH_2$) is added [72], [73] (see below).

In the scheme presented above, the NCO – NH_2 polyaddition reaction is much more rapid than the NCO – OH polyaddition reaction and requires no catalysts. This rapid curing results in short production cycles.

Prepolymers with NCO end groups can also react with water, whereby the resulting carbon dioxide byproduct is used as a foam-

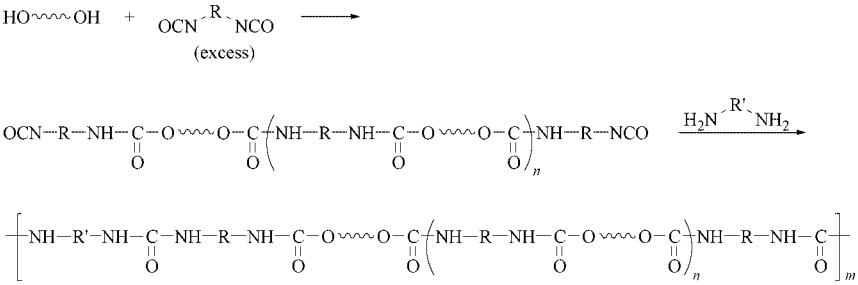

R, R′ = alkylene, arylene, heterocycle, etc.

ing agent in the production of poly(urethaneurea) foams. Poly(urethaneurea) copolymers have many practical applications as foams, cellular elastomers, fibers, aqueous dispersions, etc.

Optically active poly(urethaneurea)s were synthesized by solution polyaddition of (1*S*,2*S*)-(+)-2-amino-3-methoxy-1-phenyl-1-propanol with diisocyanates [74].

Poly(urethaneurea)s obtained from D-glucosamine and diisocyanates also have optical activity [75]. These polymers absorb 10 – 24 % water, whereas the poly(urethaneurea)s prepared from $HO(CH_2)_5NH_2$ and diisocyanates are almost non-hygroscopic; acetylation reduces the hygroscopicity.

Light-resistant poly(urethaneurea)s are prepared from 1,3-bis(isocyanatomethyl)adamantane and polyester diols with diamines as chain extenders [76].

Poly(urethaneurea)s with ionic links in the main chain were synthesized from MDI or 2,4-TDI and 4-hydroxybutyl phthalate as salts with divalent metal ions such as Ca^{2+}, Mn^{2+}, and Pb^{2+} [77].

Poly(AmideUrea)s. Aromatic poly(amideurea)s are prepared from 4-isocyanatobenzoyl chloride and diamines [78] and from 4,4′-diaminocarbanilide and $R(COCl)_2$ [79]:

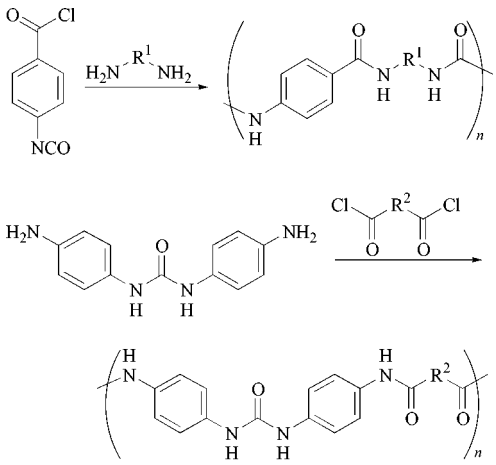

R^1, R^2 = alkylene or arylene

These poly(amideurea)s have excellent thermal stability and good solubility in aprotic dipolar solvents.

Thermostable poly(amideurea) acids can be prepared from 4,4′-diaminodiphenyl ether, pyromellitic dianhydride, and diisocyanates, which can be used as intermediates in the syntheses of poly(imideurea)s [80]:

Poly(imideurea)s with improved thermal stability are obtained from pyromellitic dian-

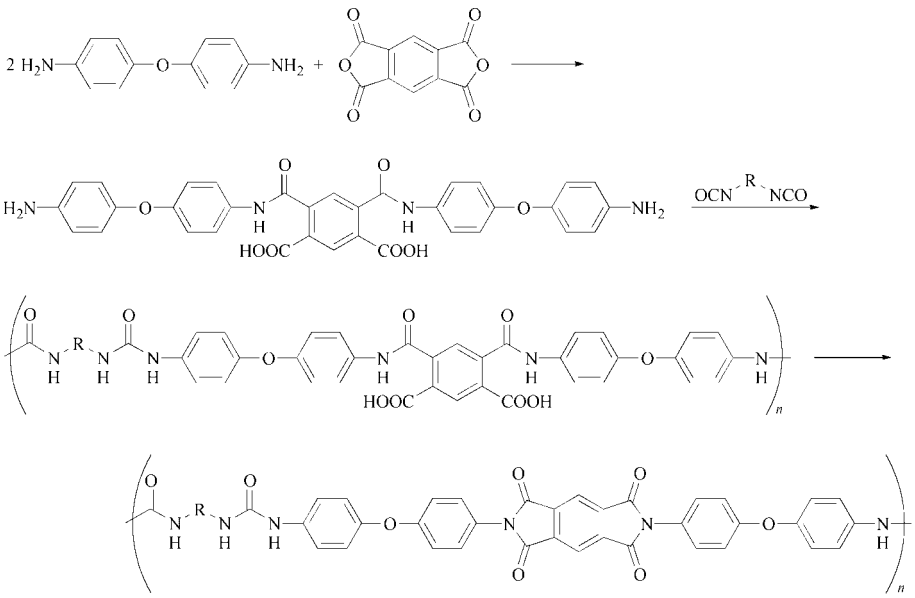

R – arylene

hydride and *N,N′*-dimethyl-*N,N′*-bis(4-aminophenyl)urea [81], or from the same dianhydride and $H_2N(NHCONHR)_mNH_2$, where R is $-(CH_2)_{12}-$; $-C_6H_4OC_6H_4-$-, $-C_6H_4CH_2C_6H_4-$-, or $-C_6H_4SO_2C_6H_4-$ [82]:

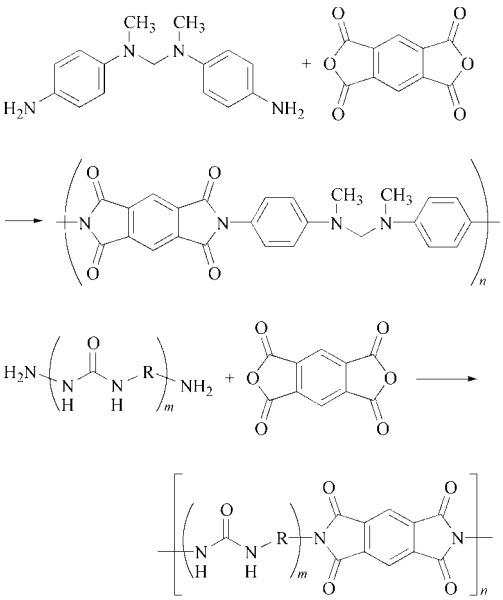

Sequential condensation can result in poly(imideurea)s with thermal stability up to 380 °C (see below).

These poly(imideurea)s can be cyclized to polyiminoquinazolinediones.

Poly(carbonateurea)s can be prepared from bis-(4-aminophenyl) carbonate and diisocyanates (a) [83], or from *p*-isocyanatobenzoyl chloride and hydroquinone, which give a monomer with NCO end groups (b) that reacts with diamines [84] (see top of the next page).

These poly(carbonateurea)s have properties similar to these of aromatic polyureas, but exhibit higher solubility.

6. Structure

The structure of polyureas and copolyureas controls their chemical and mechanical properties.

Poly(urethaneurea)s contain hard and soft segments. Usually, the hard segments have an aromatic structure and the soft segments an aliphatic structure. During solidification from solution or melt, the hard and soft segments undergo phase separation, which results in elastomeric properties. The properties of poly(urethaneurea)s depend largely on the extent of phase separation, which is due to thermodynamic incompatibility [85]. Compositional variables, such as symmetry in the diisocyanate structure [86], chain extenders, aromatic or

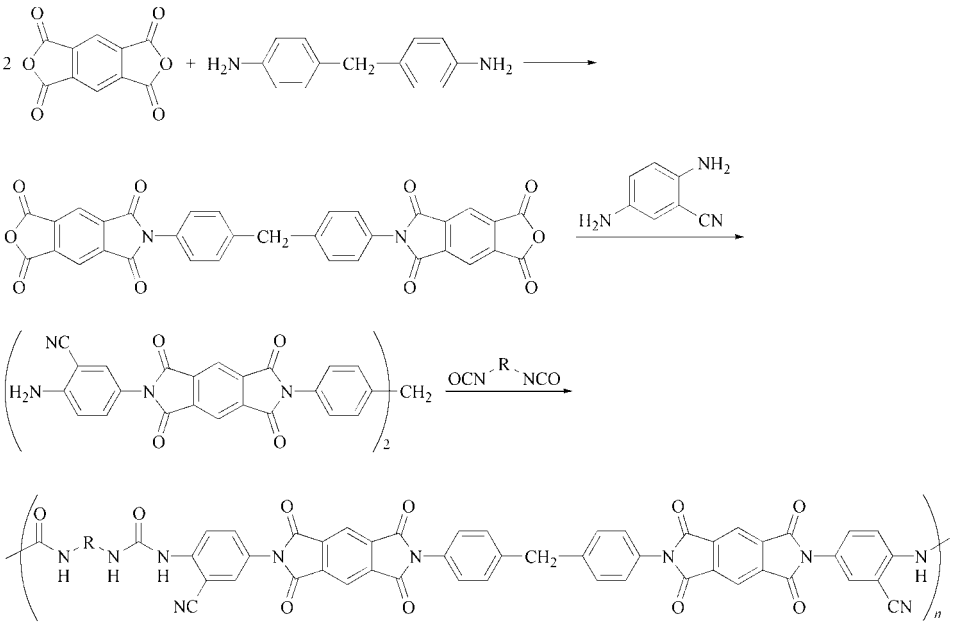

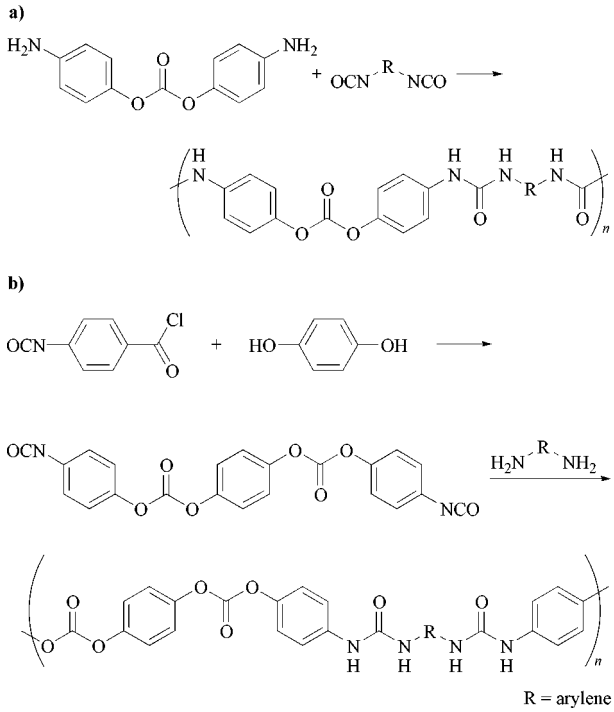

aliphatic diamines [87], molecular mass distribution of the soft and hard segments [88], cross-linking [89], and hydrogen bonding [90], have a strong influence on the extent of phase separation.

In each urethane or urea linkage, the NH and CO groups act as proton donor and proton acceptor, respectively. Since the urea group is more polar than the urethane group, the separation of hard and soft segments is more marked, the melting range of the hard segment domains is higher, and the intermolecular (interchain) interactions are stronger. The urea carbonyl groups are considered to be completely hydrogen-bonded [91]; this feature is probably responsible for the increased phase separation in poly(urethaneurea)s.

The presence of hard segments is responsible for the high tensile strength and modulus exhibited by the materials derived from these copolyureas. The soft segments determine the thermoplastic behavior.

Phase separation has been proved by electron microscopy [92], small angle X-ray scattering [93], differential scanning calorimetry [94], and other methods [95]. The same phenomenon was studied by advanced solid-state NMR spectroscopy; the sizes of various domains and the interfaces between them were quantified by spin measurements. Also, the impact of annealing, method of polymerization, and hard segment content on the properties of poly(urethaneurea)s were studied. Major tranformations have been observed after annealing at 190 °C, when the system behavior changes from "soft in hard" to "hard in soft". The highest microphase separation is observed in solution-polymerized samples. These NMR experiments reveal the nonequilibrium nature of RIM systems for poly(urethaneurea)s [96]. For the same system of RIM poly(urethaneurea)s, FT-IR spectrophotometric data suggested that an increase in degree of phase separation and a decrease in the size of hard-segment domains take place with increasing catalyst concentration [97].

Problems of structure arise in the field of polyureas and poly(urethaneurea)s expanded by water for foams. The reaction of isocyanate groups with water to give urea groups can be monitored by IR spectroscopy at 1646 cm^{-1} for carbonyl groups and 3300 – 3320 cm^{-1} for

NH groups [98]. The absorption of the carbonyl group from polyureas is not restricted to the region around 1640 cm^{-1} and varies within wide limits according to the degree of association [99]. A band at 1695 cm^{-1} has been attributed to an unassociated urea group in polyureas [100]. The NH absorption at 3320 cm^{-1} indicates that most of the NH groups are hydrogen-bonded, because the free-NH absorption at 3445 cm^{-1} is negligible [91]. This polar structure can be represented by the following structures:

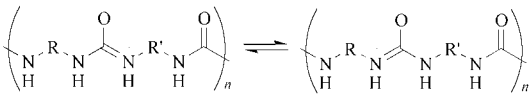

R, R' = arylene, alkylene, etc.

The high melting points, thermostability, and crystallinity of polyureas may be due to the partial polar structure.

The capacity for hydrogen bonding between urea groups and solvents is as follows: alcohols $\gg$ THF > $CHCl_3$ > toluene [101].

7. Properties

7.1. Physical Properties

Melting Temperature. The polar structure of polyureas results in high melting points that are higher than those of the corresponding polyimides. Aliphatic polyureas melt at lower temperatures than aromatic polyureas. The melting temperature of aliphatic polyureas decreases with increasing number of methylene groups. Thus, nonamethylenediamine gives polyureas with melting temperatures of about 200 – 210 °C that can be injection-molded and extruded [102]. Alkylation or acylation of the nitrogen atoms lowers the melting points [103]. The copolyureas also melt at lower temperatures. For homopolyureas, melting point variations can be seen in Table 3.

The glass transition temperature T_g also depends on structure. Aliphatic polyureas usually have a T_g below 100 °C, whereas the T_g of aromatic polyureas is above 100 °C.

Solubility. Polyureas having polar structures, are difficult to dissolve in organic solvents. Aliphatic polyureas are soluble in phenol, cresols, acetic acid, and similar solvents. Aromatic polyureas are usually soluble only in aprotic dipolar solvents such as DMF, DMA, and NMP [109]. The symmetry or asymmetry of the aromatic monomers influences the solubility of the polyureas. Copolyureas, particularly aliphatic ones, are usually more soluble than polyureas. The solubility of poly(urethaneurea)s can be improved by reaction with trifluoroacetic anhydride to give polymers with =N–$COCF_3$ groups.

Water Absorption and Permeability. The water absorption of polyureas is low and depends on structure. Water absorption is increased by incorporation of heteroatoms, such as oxygen, sulfur, –COOH, or –SO_3H, whereas it decreases with increasing number of carbon atoms. Poly(urethaneurea)s derived from D-glucosamine and diisocyanates have a water absorption of 10 – 24 % [110]. The water permeability of polyurea films is higher than that of polyamide films.

7.2. Chemical Properties

Thermal Stability. Polyureas have a thermal stability higher than that of polyurethanes

Table 3. Melting points of homopolyureas

Main chain	*mp*, °C	Prepared from	Ref.
Aliphatic	180 – 230	diamines and urea	[104]
Aromatic	260 – 320	diamines and carbon dioxide	[105]
Aliphatic/aromatic	200 – 240	aromatic/aliphatic dihalide and alkali metal cyanate	[106]
Naphthalenediyl	350 – 380	diamines and urea	[107]
Heterocyclic	330 – 350	diamines and diisocyanates	[108]

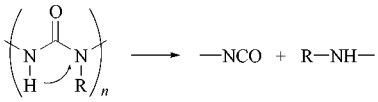

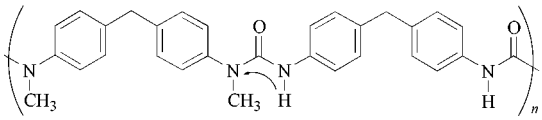

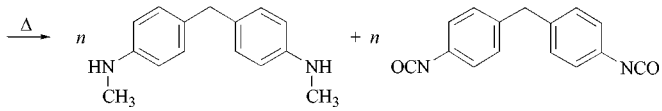

but lower than that of polyamides [111]. The thermal stability of the urea group is a complex issue. Thus, polyhexamethyleneurea ($mp = 300\,°C$) can be heated to 260 °C without decomposition, whereas polydecamethyleneurea ($mp = 230\,°C$) decomposes rapidly at 260 °C in the molten state. The mechanism of this thermal decomposition has been studied by mass spectrometry and involves a hydrogen transfer to give isocyanates and amines [112] (see above).

N,N′-Disubstituted polyureas decompose at higher temperatures than unsubstituted polyureas because of the higher energy required for the C–H hydrogen transfer, which involves two competing decompositions [112]:

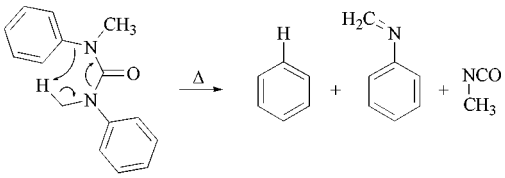

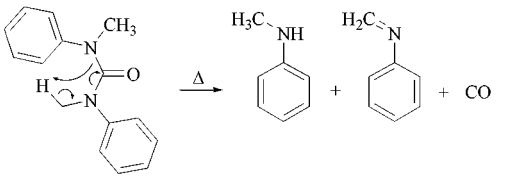

This six-centered transition state is indicated by a peak in the MS (mass spectrometry) spectrum that corresponds to methyl isocyanate [112]. The thermal decomposition of polymers containing a carbonyl group flanked by 1,4-piperazinediyl units involves hydrogen transfer from a methylene group to nitrogen [113]:

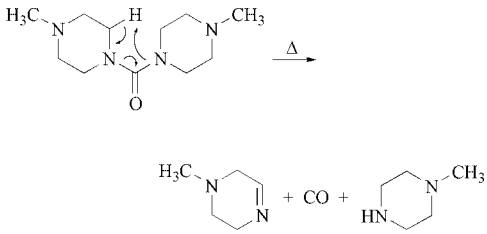

A method has been reported for the synthesis of isocyanates by thermal decomposition of polyureas that involves this interesting reaction [114].

Thermal stability of polyureas can be improved by copolymerization to give poly(phosphorylurea)s, poly(silylurea)s, poly(amideurea)s, poly(imideurea)s, etc. Thermal stabilities of *N*-substituted polyureas are higher than those of the corresponding unsubstituted polymers.

Hydrolytic Stability. Polyureas have excellent stability to hydrolysis under acidic and alkaline conditions. For example, polynonamethyleneurea fibers are more resistant to acid than nylon 6 or nylon 6,6; resistance to alkalis is higher than that of poly(ethylene terephthalate). In 40 % sulfuric acid or 40 % caustic soda solution at room temperature, these fibers lose less than 10 % of their strength after 10 h [115].

Improved resistance toward acids and bases has been reported for *N*-substituted polyureas prepared from *N,N′*-dialkylhexamethylenediamines and phosgene [116].

Biodegradation. Copolyureas containing amino acids or peptides in their chains are degraded by living organisms. Copolyureas containing dipeptides are degraded twice as fast as than those containing amino acids [117]. Similar polyureas are also degraded by chymotrypsin or subtilisin [118].

8. Uses

The urea linkages in poly(urethaneurea)s or polyureas extended with diamines cause an increase in phase separation compared to polyurethanes extended with diols. This phase separation improves various mechanical properties, such as elongation at break, hysteresis, and stress elongation [119].

8.1. Foams

Polyurethane and poly(urethaneurea) foams are generally divided into flexible, semiflexible, and rigid products, which can be produced by one-shot or by prepolymer methods.

In the one-shot method, the components are mixed directly, generally with simultaneous addition of auxiliaries such as catalysts, reinforcing agents, fillers, foam stabilizers, and flame retardants. These systems are exothermic.

In the prepolymer method, NCO prepolymers can be obtained from di- or polyhydroxy compounds, usually hydroxy-terminated polyethers or polyesters, which react with di- or polyisocyanates in excess. These NCO prepolymers are important compounds, because they can be cured with active hydrogen atoms from diamines, diols, water, etc. Many polyurethane and poly(urethaneurea) elastomers are produced via NCO prepolymer intermediates.

The curing of NCO prepolymers with diols is used in the production of polyurethane elastomers. The same reaction, but with diamines, gives poly(urethaneurea) elastomers. With aliphatic amines (strong nucleophiles), this reaction is too fast to allow satisfactory mixing and a sufficiently long pot life. For this reason, less reactive aromatic diamines are used as chain extenders, in particular sterically hindered diamines or diamines whose nucleophilicity is weakened by electron-withdrawing groups.

Chain extension with diamines has advantages. For example, this curing process occurs rapidly and without additional heating; short production cycles are therefore possible. Usually, the NCO/NH_2 reaction ratio is about 1 : 1. Symmetrical diamines with deactivating substituents are used for elastomer synthesis, and alkyl-substituted toluenediamines, such as diethyltoluenediamine, are usually employed for reactions with NCO prepolymers in the RIM process.

If this reaction is carried out in polar organic solvents, practically any diamine can be used, even especially reactive aliphatic ones such as ethylenediamine.

The NCO/H_2O reaction (chemical blowing process) is a very important process for the production of foams. This chemical blowing process can be supplemented by physical blowing with low-boiling liquids as blowing agents, such as trichlorofluoromethane, which can be used to reduce the foam density further.

Organometallic compounds such as organotin compounds, in combination with tertiary amines, have a synergistic effect in the catalysis of the isocyanate – water reaction (Table 4) [120].

Linear or slightly branched polyols with a relatively high molecular mass (2000 – 8000) yield flexible foams. Combinations of these polyols with low molecular mass glycols or amines as chain extenders afford semiflexible foams. Highly branched polyols with a relatively low molecular mass (< 1000) yield rigid foams.

Polyurea foam materials are also prepared from diisocyanates with different structures by reaction with water. For flexible foams, linear or slightly branched prepolymers are used; more highly branched prepolymers produce rigid polyurea foams.

Flexible foams are prepared from diisocyanates, polyfunctional polyols, blowing agents, catalysts, and surfactants. Although various diisocyanates have been used for flexible foams, toluene-2,4-diisocyanate (TDI) is preferred. Modifiers, fillers, and plasticizers are used to confer special properties. The flexibility of these foams depends on the molecular mass of the polyesters and polyethers used. The ether group

Table 4. Effects of catalysts on relative reactivities of phenyl isocyanate with active hydrogen compounds in dioxane at 70 °C

Catalyst,	Relative reactivity with		
10 mol %	*n*-Butanol	Water	Diphenylurea
None	1.0	1.1	2.2
N-Methylmorpholine	40	25	10
Triethylamine	80	47	4
Tetramethyl-1,3-butanediamine	260	100	12
Triethylenediamine	1.20	380	90
Tributyltin acetate	80	14	8
Dibutyltin diacetate	600	100	12

imparts flexibility, which, in turn, promotes softness, low melting point, low glass transition temperature, and elasticity. Surfactants promote intimate mixing of the components, regulate and control cell size, and impart stability. Many types are used, such as organic nonionic or ionic surfactants and silicones.

The properties of flexible foams also depend on the processing conditions (mixer speed, air-injection rate, etc.). Polyester-based flexible foams have higher tensile strength and elongation-at-break than polyether foams.

Flexible foams are characterized by reversible deformability and open-cell structure that is associated with air permeability. They have many applications, for example, mattress manufacture, automobiles, and aircraft construction.

Semiflexible foams are used almost exclusively as moldings and are 90 % open-celled. They have lower elasticity and good damping properties. In consequence, they are used as protective cushioning for control panels and consoles and for door and pillar paneling.

Rigid foams are prepared similarly to flexible foams but with different starting materials. In rigid foams, groups that restrict chain rotation, particularly aromatic rings, confer stiffness, which increases the melting point and hardness, and reduces elasticity.

Polyether resins used for rigid foams have lower molecular mass (300 – 1000) than those for flexible foams (3000 – 5000). Polyethers with aliphatic/aromatic structure, prepared from aromatic diamines and propylene oxide, contribute to the stiffness of rigid poly(urethaneurea) foams.

The production of rigid poly(urethaneurea) foams from hydroxy-terminated resins and isocyanates requires auxiliaries such as blowing agents, catalysts, cell-size regulators, flame retardants, surfactants, and cross-linking agents. As with flexible foams, the reaction between diisocyanates and water generates CO_2, which serves as blowing agent. A halogenated hydrocarbon, such as trichlorofluoromethane, may also be employed, because this inert liquid with *bp* 23.8 °C is quickly volatilized by heating. The gas bubbles expand the foam and remain in the closed cells of the final product. In addition, CCl_3F has a low thermal conductivity and is nonflammable.

The density of rigid foams depends on the amount of gas trapped in the cells. A commercial rigid foam that contains 3 % solid polymer and 97 % gas has a density of about 16 kg/m^3. These foams are manufactured in such a way as to minimize convection currents [121]. Rigid foams are predominantly of the closed-cell type. For special uses, open-cell types are also produced.

Mechanical properties depend on the composition used in the formulation and the plastic nature of the foam at high temperature. Rigid poly(urethaneurea) foams provide excellent thermal insulation, especially when blown with fluorocarbons; high strength combined with light weight; heat resistance; and adhesion to wood, glass, ceramics, and fibers. The rigid foams can be used between − 200 and + 130 °C, and for short periods they can tolerate + 250 °C. As cross-linked plastics, the rigid foams are largely stable to chemicals, such as dilute acids and alkalis. The chief uses of rigid foams are in the heat and cold insulation field. They are used for insulation of refrigerators, in-

cluding industrial refrigerators, freezers, and hot water accumulators, and as insulation for industrial tanks, containers, etc.

Polyurea foams are prepared by the reaction of diisocyanates and water; the mechanism is similar to that involved in the preparation of poly(urethaneurea) foams. In this reaction, tertiary amines are used as catalysts. Cu(II) and Zn(II) salts and tertiary amines exhibit a synergistic effect in the catalysis of the MDI – water reaction, which gives polyurea foams with satisfactory properties [122]. In these formulations, when *N,N′*-dimethylaminoethanol is used as catalyst, better results are obtained with copper nitrate as cocatalyst than with zinc nitrate. Flexible polyurea foams are used for cushions and in the automotive industry, where they provide light weight, high strength, and ease of fabrication. Rigid polyurea foams are used for refrigerators and in constructions.

8.2. Moldings

The production of polyurethanes, poly(urethaneurea)s, and polyureas by reaction injection molding (RIM) is now a well-established process. It offers several advantages for high-volume production of large parts: low pressure (< 343 kPa), low temperatures (54 – 60 °C) and use of reactive liquid intermediates [123]. Since the urethane polymerization process is highly exothermic, minimal heat input is required to maintain tooling temperature during production. With polyols, isocyanates, and extenders, different polymers can be formulated with a wide range of properties, from a flexible elastomer to a rigid plastic. The final product can be solid, microcellular, or foam-modified with a variety of fillers. Glass fiber reinforced RIM (RRIM) provides dimensional stability and high impact strength [124].

Poly(urethaneurea) and polyurea elastomers represent new classes of commercial RIM elastomers. They are obtained from diisocyanates and macrodiols or macrodiamines with hindered aromatic diamines as extenders. Diethyltoluenediamine (DETDA) is used in large quantities as an extender for these elastomers which have high thermostability [125]. Because of the high rate of urea formation, aliphatic diamine extenders have found limited application in RIM.

Poly(urethaneurea) and polyurea elastomers have many practical applications; the most important is in the automotive industry. To reduce the weight of automobiles, high-modulus RIM and RRIM materials suitable for external body panels have been developed from these elastomers. In addition, these elastomers are corrosion- and impact-resistant.

Because of their mechanical and thermal properties, these rigid foams are especially suitable for structural applications and can compete with wood, glass fiber reinforced plastics, and other materials.

High-modulus poly(urethaneurea) RIM elastomers are prepared from a prepolymer of MDI and diamine, using a system termed Bayflex 110 [126]. Some properties of Bayflex 110 follow:

Density	1042 kg/m^3
Flexural modulus	
at − 30 °C	829 MPa
at 20 °C	359 MPa
at 65 °C	241 MPa
Tear strength	105 kN/m
Tensile strength	26.2 MPa
Elongation	275 %

These RIM elastomers, manufactured by Mobay Chemicals, have superior physical properties and can be released from the mold after 15 s with excellent green strength.

Polyurea RIM elastomers with high modulus can be prepared from a polyamine (Jeffamine D-2000), polyisocyanate, and an amine-terminated chain extender in an Accuratio V-R100 RIM machine, and released after 1 min. The resulting elastomer is cured at 121 °C for 1 h; properties are given below:

Tensile strength	23 – 28 MPa
Elongation	280 – 360 %
Tear strength	65 – 68 kN/m
Flexural modulus	
at 25 °C	148.2 – 148.9 MPa
at − 29 °C	526 – 599 MPa
Heat sag	6.8 – 11.5 mm

The molding behavior of the RIM polyurethane depends on the catalyst [127]. The catalyst used for the isocyanate – water reaction can affect the physical properties of the final product. Because of yellowing under the influ-

ence of UV light, the moldings must be coated with a paint of high opacity, unless they are colored black or dark brown in the solid. In the photodegradation of MDI-based poly(urethaneurea) elastomers, photo-Fries products have been observed by UV and IR spectroscopy [128].

Polyols in polyurethane and poly(urethaneurea) elastomers increase the hygroscopicity and affect the physical properties. These elastomers are less stable at elevated temperature than polyurea elastomers. Polyisocyanates react slowly with polyols, which prevents automation of the process.

Advances in RIM technology include the development of new polyurea systems for the production of automobile body panels. These catalyst-free systems are based on amine-terminated polyesters. In automobile body panels, polyurea RIM materials offer greater impact resistance at equivalent glass-flake reinforcement; improvement in thermal stability for less distortion during paint cure cycles; and lower water absorption for better environmental resistance. These polyureas are usually prepared by the reaction of an amine – MDI-based isocyanate with an amine-terminated polyether (instead of the standard hydroxy-terminated polyether) and an aromatic diamine as chain extender.

Polyurea RIM elastomers present the following advantages:

- The autocatalytic reaction of isocyanate with amine is faster than the isocyanate – water reaction and results in high modulus polymers for automotive applications
- Catalysts are not necessary, and the process can be automated
- Decomposition and yellowing of the final product do not occur
- The higher thermostability extends their practical utility

Table 5 shows the development of RIM technology for polyurethane, poly(urethaneurea), and polyurea [129]. Texaco produces a polyether with amino end groups and a molecular mass of ca. 4000 (Jeffamine), which can be used to replace the polyols in polyurethanes and poly(urethaneurea)s. Mobay Chemicals investigated and established the optimum conditions for poly(urethaneurea) systems (Bayflex 110) and polyurea systems (Bayflex 150). Dow Chemicals investigated Spectrim HT systems; ICI developed the polyurethane Rubicon; and Admiral Equipment investigated machines with a high degree of automation adapted for a very fast RIM process.

The PHD (Polyharnstoff Dispersion) polyols constitute an important class of poly(urethaneurea)s with applications such as RIM elastomers, which were first prepared at Bayer by a polyaddition reaction in situ from diisocyanates and diamines in a polyol medium [130].

8.3. Fibers

Polyureas can be used for fibers because they have good physical and mechanical properties, melting points over 200 °C, low specific gravity, high stability toward acids and alkalis, and excellent dyeability. For example, polynonamethyleneurea has a melting point of 237 °C and a softening point of about 205 – 210 °C, which are suitable properties for fibers.

Polyureas with fiber-forming properties are also obtained by the reaction of $H_2N(CH_2)_nNH_2$ and $O[(CH_2)_4NCO]_2$. For $n = 4$, the resulting polyurea melts at 223 – 225 °C and is suitable for fibers [131].

A copolymer polyester–polycarbonate – polyurea is suitable for spandex fibers with good resistance to environmental degradation [132].

In an interesting continous solventless process, elastomeric filaments and fibers with a poly(urethaneurea) structure have been prepared by spinning in the gas phase from the corresponding monomers [133].

The addition of a polyurea to polyoxymethylene gives heat-stable transparent fibers with good dye affinity [134].

A polyester – cotton blend with improved luster and washfastness was prepared by impregnating the fabric with a diamine and then with a diisocyanate; calendering the fabric formed the polyurea in situ [135].

8.4. Films and Membranes

An interesting application of polyureas is for new nonlinear optical materials prepared as thin films from aromatic polyureas. For example,

Table 5. Development of polyurethane, poly(urethaneurea) and polyurea RIM elastomers

Parameter	Polyurethanes	Poly(urethaneurea)s	Polyureas (1985)	
	(1974 – 1975)	(1980 – 1984)	at 20 °C	at 70 – 75 °C
Productivity, kg/s	2.3 – 3.2	3.2 – 4.5	5.4 – 6.8	10.9 – 13.6
Cream time, s	4 – 8	2.5 – 3	2 – 2.5	1 – 1.5
Total production cycle∗, s	180 – 420	180	60 – 70	60

∗ Rise time.

such thin films were prepared by simultaneous deposition of 4,4′-diaminodiphenylmethane and MDI. Investigations suggest that the poling process of polymerization of oligomers takes place at about 116 °C under a high electric field to give polar urea bonds with a large residual polarization. This polarization determines the pyroelectic and piezoelectric activities after poling treatment [136].

Similar results have been obtained for polyureas prepared from α,α'-dianilino-*p*-xylene and MDI, whose large second-order optical nonlinearity coupled with their unique optical transparency in the UV region make them potentially useful materials for nonlinear optics [137].

Microporous films for artificial leather can be obtained from poly(urethaneurea)s, provided these materials have adequate mechanical properties and good permeability to water [138]. Large quantities of these polymers are used for artificial leather.

Poly(urethaneurea) membranes with high H_2S/CH_4 selectivity have been obtained [139].

Ultrathin polyurea membranes are employed for water desalination by reverse osmosis. The urea group in polyureas is hydrophilic and resistant to chemicals and organic solvents; thus, these polymers are suitable for ultrafiltration applications [140].

Polyurea films are useful as flexible materials impermeable to air and other gases and for construction of inflatable boats, balloons, and similar applications [141].

Polyisocyanates with the general formula $R(CH_2)_y(NCO)_x$, in which R = alkyl, $y = 0$ or 1, and $x = 2 - 4$, react with polyamines; the resulting polyurea films have good adhesion to steel and aluminum and high resistance to chemicals and solvents [142].

Semipermeable membranes with high salt rejection have been prepared from polyureas [143].

The polyurea films produced by an interfacial polymerization reaction between water and a triisocyanate facilitate the formation of stable emulsions [144].

8.5. Coatings

Coating materials can be made from polyureas or poly(urethaneurea)s in solution or emulsion. A new spraying technology has many practical applications; polymers thus applied show excellent adhesion to a variety of substrates, including sand-blasted or primed steel, aluminum, and concrete. The products include two-component sprayed polyurea elastomers, prepared from amine-terminated polyether resins and polyisocyanates. The polyurea spray elastomers systems require no catalyst and have extremely high reactivity and cure [145]. In other applications, aqueous polyurea dispersions can be used for coatings with improved hardness and solvent resistance [146].

8.6. Biomedical Applications

Several polyurea and poly(urethaneurea) materials exhibit blood compatibility and have found applications in medical equipment and artificial organs [147]. The polymer membranes in pneumatic blood pumps [148] should have high tensile strength, low permanent deformation, low hydrolyzability, and high purity.

The hydrolytic stability of poly(urethaneurea)s in extracellular solutions was studied. After 50 d, failure was observed. After 120 d, resistance decreased by 15 – 20 % due to hydrolytic degradation of the polymers. The rate

of hydrolysis was higher in extracellular solutions than in water [149]. Poly(esterurethaneurea)s are not suitable for medical applications because of their hydrolytic instability caused by the ester group.

Cross-linkable polyureas have been prepared for contact lenses [150].

8.7. Other Applications

Polyureas are used as thickeners for lubricating oils; the resulting lubricating greases have long lifetime at high temperatures, high oxidation resistance, good corrosion protection, high structural stability, and good low-temperature properties [151].

Polyureas and poly(urethaneurea)s can be used for the microcapsulation of herbicides, inks, colorants, medicines, and other products [152], [153].

Polyureas with carboxylic or sulfonic groups are good ion exchangers [154].

Polyureas are used for the removal of oils and bitumen from wastewater by filtration [155].

Petroleum production from oil wells is increased by injection into the borehole of a polyamine followed by a polyisocyanate. These react in situ and form polyureas that contribute to the consolidation of the oil sand and prevent clay swelling [156].

Trimellitic anhydride, polyisocyanates, and polyamines give electric insulator compositions for Cu wire with excellent flexibility and resistance to heat, abrasion, solvents, and thermal shock [157].

9. Safety and Environmental Aspects

Due to the wide range of applications of polyureas, copolyureas, and especially poly(urethaneurea)s, millions of people come into contact with them.

Many years' experience in handling the raw and auxiliary materials for these polymers has shown that these products can be handled safely. Air extraction systems must be present at all workplaces where TDI is processed at room temperature or MDI at temperatures above 40 °C. Aminopolyethers are moderately irritating to the eyes and skin; hence, contact with them should be avoided. The same applies to tertiary amines used as catalysts.

Since the carbon dioxide formed in the reaction of water with isocyanates is a natural constituent of air, it presents no problem. The other blowing agents are discharged into the outside air.

Wastes for which no utilization is possible can be disposed of in incineration plants. Granulated RIM wastes can be molded by the flow molding process at high pressure and high temperature to give new moldings. Fine-ground wastes can be used as fillers in production.

The glycolysis of reinforced reaction injection moldings (RRIM) of these polymers with diethylene glycol or dipropylene glycol at high temperatures can be used for the recovery of polyols and aromatic polyamines as suitable products for new RIM [158].

Organic isocyanates can be prepared by thermal decomposition of polyureas at 220 °C [114]. Another route is to utilize the energy by incineration of these polymers, but the calorific yield is small in comparison with other polymers, for example, polyethylene.

10. References

Specific References

1. Institut de Petrole, des Carburants et Lubrifiants, DE-O 1 946 942, 1968 (J. C. Silion, G. B. Gaudemaris); *Chem. Abstr.* **72** (1970) 123035j.
2. DuPont, US 2 888 438, 1959; 2 975 157, 1961 (M. Katz).
3. S. V. Joshi, A. Rao, *J. Appl. Polym. Sci.* **28** (1983) 1457.
4. *Mater. Plast. Elastomeri,* (1984) April, 217.
5. *Eur. Chem. News* (1985) Oct., 21.
6. J. T. Davis, F. Ebersole, *J. Am. Chem. Soc.* **56** (1934) 885.
7. R. Scholl, K. Holdermann, *Ann.* **345** (1907) 376.
8. R. Craven, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **16** (1956) no. 2, 3.
9. J. Baker, D. N. Bailey, *J. Chem. Soc.* (1957) 4649.
10. C. A. Naegeli, L. Conrad, *Helv. Chim. Acta* **21** (1938) 1127.
11. BASF, EP-A 896 010, 1999 (D. T. Ulrich).
12. Y. Watabe, M. Ishi, Y. Iseda, *J. Appl. Polym. Sci.* **25** (1980) 2747.

13. E. V. Scriven, *J. Chem. Soc. Rev.* **12** (1983) 136.
14. R. B. Moodie, *J. Chem. Soc. Perkin Trans.* **2** (1981) 664.
15. S. I. Axelrood, C. W. Hamilton, *Ind. Eng. Chem.* **53** (1961) 889.
16. F. Millich, C. Carraher: *Interfacial Synthesis,* **vol. 2,** Marcel Dekker, New York 1977.
17. G. S. Kumar, *Macromolecules* **17** (1984) 2463.
18. O. Bayer, *Angew. Chem.* **59** (1947) 263.
19. D. J. Williams, *Angew. Chem., Int. Ed.* **23** (1984) 690.
20. D. P. Loker, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **39** (1998) 579.
21. A. Ulman, D. J. Williams, C. S. Willand, *J. Am. Chem. Soc.* **112** (1990) 7083.
22. N. Nishi, S. Tokura, *Makromol. Chem.* **192** (1991) 1881.
23. A. Kumar, E. W. Meijer, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **39** (1998) no. 2, 619.
24. C. J. Hawker, P. Farrington, *J. Am. Chem. Soc.* **117** (1995) 6123.
25. J. A. Saunders, K. C. Frisch: *Polyurethane, Chemistry and Technology, Part 1: Chemistry,* Wiley Interscience, New York 1962.
26. P. Merkaert, R. Jerome, *J. Appl. Polym. Sci.* **27** (1982) 4221.
27. J. M. Borsus, Ph. Tayssie, *J. Appl. Polym. Sci.* **27** (1982) 4029.
28. N. Yamazaki, T. Iguchi, F. Higashi, *J. Polym. Sci., Chem. Ed.* **13** (1975) 785.
29. C. I. Chiriac, *Polym. Bull.* **15** (1986) 65.
30. T. Lieser, H. Gehlen, *Liebigs Ann. Chem.* **556** (1944) 127.
31. DuPont, US 2 816 879, 1957 (E. L. Wittbecker).
32. H. Iiyama, M. Asakura, K. Kimoto, *Kogyo Kagaku Zasshi* **68** (1966) 236.
33. G. S. Kolesnikov, *Vysokomol. Soedin., Ser. A* **11** (1969) 2682.
34. N. Yamazaki, F. Higashi, *J. Polym. Sci., Chem. Ed.* **17** (1979) 835.
35. DuPont, US 2 190 770, 1940 (W. H. Carothers).
36. SU 905 228, 1980 (R. D. Katsarava); *Chem. Abstr.* **97** (1982) 24396n.
37. A. M. Schmuki, W. U. Suter, *Macromolecules* **18** (1985) 823.
38. W. R. Grace, FR 1 299 698, 1962.
39. N. Ogata, M. Watanabe, *J. Polym. Sci., Polym. Lett. Ed.* **24** (1986) 66.
40. M. Srinivasan, *Proc. Matl. Symp. Therm. Anal.* **9** (1993) 167; *Chem. Abstr.* **122** (1995) 315230s.
41. L. Takarsewski, J. Ossowski, W. Pietranek, *Polimery (Warsaw)* **39** (1994) 738; *Chem. Abstr.* **123** (1995) 56774f.
42. S. Ayusman, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **33** (1992) no. 2, 180; *Chem. Abstr.* **120** (1994) 31397k.
43. S. Kanamaru, T. Takata, *Macromolecules* **27** (1994) 7492.
44. G. Caraculacu, E. Scortanu, A. Caraculacu, *Eur. Polym. J.* **19** (1983) 143; **34** (1998) 1265.
45. Y. Imai, *J. Polym. Sci., Part A-1* **5** (1967) 2289.
46. Bayer, EP-A 743 332, 1996 (J. E. Haakan); *Chem. Abstr.* **126** (1997) 47677m.
47. Toyo Rayon, JP 69 19 556, 1969 (N. Yoda, M. Kurihara).
48. E. Dyer, *J. Polym. Sci., Part A-1* **7** (1969) 833.
49. Asai Shokusan Y. K., JP 11 106 507 20, 1999 (H. Okamoto); *Chem. Abstr.* **130** (1999) 313020q.
50. Mitsui Toatsu Chemicals, JP 08 259 603, 1996 (M. Murata); *Chem. Abstr.* **126** (1997) 48436u.
51. DuPont, US 16 475, 1997 (M. Teasley); *Chem. Abstr.* **127** (1997) 34716u.
52. SU 590 962, 1973 (N. P. Smetankina, N. N. Laskovenko); *Chem. Abstr.* **95** (1981) 8239s.
53. N. Laskovenko, *Vysokomol. Soedin.* **66** (1993) 203; *Chem. Abstr.* **119** (1993) 181375v.
54. C. E. Carraher, C. W. Krueger, *Makromol. Chem.* **133** (1970) 219.
55. J. A. Mikroyannidis, *J. Polym. Sci., Polym. Chem. Ed.* **22** (1984) 3423.
56. Texaco Belgium, GB 1 514 511, 1978 (C. J. Hermans).
57. Aerojet General Corporation, US 5 637 772, 1997 (A. Malik, P. Carlson); *Chem. Abstr.* **127** (1997) 95718t.
58. T. Lesiak, T. Nowakovski, *J. Prakt. Chem.* **321** (1979) 921.
59. G. S. Goldin, *Vysokomol. Soedin., Ser. B* **12** (1970) 307.
60. Toyo Rayon, JP 69 19 555, 1968 (N. Yoda, M. Kurihara).
61. T. Yasugawa, H. Yamaguchi, *J. Polym. Sci., Polym. Chem. Ed.* **17** (1979) 3387.
62. V. V. Korshak et al., *Vysokomol. Soedin., Ser. A* **22** (1980) 41.
63. S. J. Huang, V. K. Leong, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **20** (1979) 552.
64. R. D. Katsarava, *Makromol. Chem.* **194** (1993) 3209.
65. N. N. Barashkov, E. N. Teleshov, *Vysokomol. Soedin., Ser. A.* **20** (1978) 2749.
66. A. P. Grecov, K. H. Kornev, *Khim. Geterotsikl. Soedin.* **5** (1967) 806; *Chem. Abstr.* **68** (1968) 22287e.

67. F. A. Neugebauer, H. Trischmann, *J. Polym. Sci., Part B* **6** (1968) 255.
68. T. Yuki, S. Sakurai, T. Noguchi, *Bull. Chem. Soc. Jpn.* **43** (1970) 2130.
69. SU 569 584, 1977 (V. V. Korshak, N. I. Bekasova); *Chem. Abstr.* **87** (1977) 185488e.
70. K. Gonsalves, D. Marvin, *J. Am. Chem. Soc.* **106** (1984) 3862.
71. Nippon Soda Co., Ltd., JP Kokai 77 109 593, 1977 (H. Okamura).
72. DE 120 107, 1976 (K. Walter, H. Mey); *Chem. Abstr.* **86** (1977) 73692n.
73. R. S. Ward, *J. Appl. Polym. Sci.* **27** (1982) 2167.
74. C. Jun, *J. Polym. Sci., Part A: Polym. Chem.* **31** (1993) 1719; *Chem. Abstr.* **119** (1993) 117957j.
75. Y. Iwakura, *Makromol. Chem.* **178** (1977) 2939; **180** (1979) 2331.
76. SU 647 313, 1979 (A. P. Khardin, N. G. Gureev); *Chem. Abstr.* **90** (1979) 205549r.
77. R. A. Prasath, *Macromolecules* **31** (1998) 110.
78. K. Hayashi, S. Hang, *Makromol. Chem.* **86** (1965) 64.
79. Firestone Tire and Rubber Co., US 4 035 344, 1977 (W. J. Spiewak).
80. V. P. Telesov, E. N. Provednikov, *Vysokomol. Soedin., Ser. A* **20** (1978) 1586.
81. Y. Imai, K. Masaaki, *Makromol. Chem. Phys.* **199** (1998) 457.
82. Société Rhodiaceta, DE-O 1 949 281, 1970 (K. Grundschober, C. Carvini); *Chem. Abstr.* **73** (1970) 15534h.
83. O. V. Smirnov, *Vysokomol. Soedin., Ser. A* **11** (1961) 2211; *Chem. Abstr.* **72** (1970) 32287b.
84. Mitsui Toatsu Chemicals, Inc., JP Kokai Tokyo Koho JP 82 40 452, 1982.
85. C. S. Paik Sung et al., *Macromolecules* **12** (1979) 538.
86. C. S. Paik Sung: *Polymer Alloys,* Plenum Press, New York 1977, p. 261.
87. C. S. Wu et al., *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **19** (1978) no. 2, 679.
88. R. W. Seymour, S. C. Cooper, *Polymer* **14** (1973) 255.
89. V. V. Shilov, *J. Appl. Polym. Sci.* **29** (1984) 1912.
90. J. L. Blackwell, K. H. Gardner, *Polymer* **20** (1979) 13.
91. M. Shibayama et al., *Polym. J.* **18** (1986) 719.
92. C. Dequartre, J. P. Pascault, *Angew. Makromol. Chem.* **72** (1978) 11.
93. S. L. Samuels, G. L. Wilkes, *J. Polym. Sci., Part C* **43** (1973) 149.
94. Y. Camberlin, *J. Polym. Sci., Polym. Chem. Ed.* **21** (1983) 415.
95. G. Wilkes: *Adv. Chem.,* 1979, Ser. **176** (Multiphase Polym.) 53; *Chem. Abstr.* **90** (1979) 204991s.
96. S. Lehmann, A. Donald, *J. Polym. Sci., Part B* **36** (1998) 693.
97. L. Ming, S. Kahn, *J. Polym. Sci., Part B* **35** (1997) 865.
98. F. E. Bailey, *Proc. IUPAC, Macromol. Symp.* **28** (1982) 100.
99. J. Hocker, *J. Appl. Polym. Sci.* **25** (1980) 2879.
100. K. Saito, *J. Macromol. Sci. Phys.* **10** (1974) 591.
101. L. Ming, *Polym. Mater. Sci.* **77** (1992) 562; *Chem. Abstr.* **127** (1997) 234890h.
102. Schering, NL-A 6 614 989, 1967; *Chem. Abstr.* **67** (1967) 65023n.
103. S. Caruso et al., *J. Polym. Sci., Polym. Chem. Ed.* **20** (1982) 1690.
104. P. Bornev, R. Paseday, *Makromol. Chem.* **101** (1967) 1.
105. N. Yamazaki, S. Nakahama, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **20** (1979) no. 1, 146.
106. Marathon Oil, FR 1 545 708, 1968 (A. P. Aragabright, J. V. Sinkey).
107. G. S. Kolesnikov, *Vysokomol. Soedin., Ser. B* **9** (1967) 242.
108. V. P. Telesov, *Vysokomol. Soedin., Ser. A* **20** (1978) 1586.
109. F. Higashi, *Polym. Prepr. Jpn.* **30** (1981) 143.
110. K. Kurita, *Kenkya Hokoku-Asahi* **37** (1980) 59; *Chem. Abstr.* **94** (1981) 209259b.
111. M. Y. Ah, *Han'guk Somyu Konghakhoechi* **23** (1986) 130; *Chem. Abstr.* **105** (1986) 153623s.
112. G. Montando, D. Vitalini, *J. Polym. Sci., Polym. Chem. Ed.* **21** (1983) 3321.
113. S. Foti et al., *Anal. Chem.* **54** (1982) 671.
114. PCT Int. Appl. WO 98 54 129, 1998 (J. Bosman); *Chem. Abstr.* **130** (1999) 14324y.
115. R. W. Moncrieff: *Man-Made Fibers,* John Wiley & Sons Inc., New York 1963, p. 356.
116. Union Carbide, US 3 130 170, 1964 (R. J. Cotter).
117. T. E. Lipatova, *Vysokomol. Soedin., Ser. B* **26** (1984) 149.
118. S. J. Huang, K. W. Leong, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **20** (1979) no. 2, 552.
119. T. V. Smith, N. H. Sung, *Macromolecules* **13** (1980) 121.
120. K. A. Pigott: *Encyclopedia of Polymer Science and Technology,* **vol. 11,** John Wiley & Sons, Inc., New York 1969, p. 524.

121. R. H. Harding, B. F. James, *Modern Plast.* **39** (1962) 133.
122. J. M. Borsus, R. Jerome, *J. Appl. Polym. Sci.* **26** (1981) 3027.
123. G. D. Lewis, *J. Elastomers Plast.* **14** (1982) 3.
124. S. J. Monte, G. Sugerman, *J. Elastomers Plast.* **14** (1982) 34.
125. R. J. G. Dominguez, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **25** (1984) no. 2, 293.
126. S. H. Mitzger, K. Seel, *J. Cell. Plast.* **17** (1981) 268.
127. R. E. Camargo et al., *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **25** (1984) no. 2, 294.
128. C. E. Hoyle, H. Shah, *Report* (1993), Order No. Ad-A267308; *Chem. Abstr.* **123** (1995) 289207p.
129. *Polyplasti e Plastici Rinforzati* **343** (1986) May, 47.
130. Bayer, DE 39 001, 1963 (E. Müller).
131. Asahi Chemical Industry, JP 67 5437, 1963 (K. Satome, K. Sato).
132. DuPont, WO 98 25 986, 1998 (H. Shingo, T. Matsuda); *Chem. Abstr.* **129** (1998) 82725j.
133. US 5 616 675, 1997 (K. Wilkinson); *Chem. Abstr.* **126** (1997) 306357u.
134. Kurashiki Rayon, JP 69 00 914, 1969 (K. Tanabe, K. Matsubayashi).
135. Nitto Boseki, JP Kokai Tokyo Koho 80 107 578, 1980.
136. S. Wang, *Jpn. Appl. Phys., Part 1* **32** (1993) 2768; *Chem. Abstr.* **119** (1993) 73567a.
137. T. Watanabe, *Synth. Met.* **57** (1993) 3895; *Chem. Abstr.* **120** (1994) 55555m.
138. Farbenfabriken Bayer, GB 1 167 670, 1969, (D. Dieterich, A. Reischil).
139. G. Chaterjee, S. A. Stern, *J. Membr. Sci.* **135** (1997) 99; *Chem. Abstr.* **127** (1997) 294210w.
140. P. Zschocke, H. Stratmann, *Angew. Makromol. Chem.* **73** (1978) 1.
141. Vereinigte Seidenwebereien, DE-O 2 635 114, 1978; *Chem. Abstr.* **88** (1978) 137836p.
142. General Mills, NL-A 6 607 543, 1966; *Chem. Abstr.* **66** (1967) 96307s.
143. Nitto Denko, JP Kokai Tokyo Koho 04 338 225, 1992 (J. Kitamura, A. Ikeda); *Chem. Abstr.* **119** (1993) 50909z.
144. T. Mikami, *J. Dispersion Sci. Technol.* **14** (1993) 71; *Chem. Abstr.* **118** (1993) 82038i.
145. D. Primeaux, *Int. Symp. Exhib.* (1998) 224; *Chem. Abstr.* **129** (1998) 137175n.
146. Bayer Corp., US 5 569 706, 1996 (P. B. Jacobs); *Chem. Abstr.* **125** (1996) 331744e.
147. D. J. Lyman, *J. Polym. Sci., Polym. Lett. Ed.* **18** (1980) 411.
148. H. D. Stenzenberger, D. O. Hummel, *Angew. Makromol. Chem.* **82** (1979) 103.
149. T. K. Khlystalova, M. B. Pestova, *Mech. Kompoz. Mater. Zinatne* **6** (1985) 1096; *Chem. Abstr.* **104** (1986) 136041n.
150. Novartis, WO 99 14 253, 1999 (F. Stockinger); *Chem. Abstr.* **130** (1999) 252850m.
151. H. Zajezierska, S. Patan, *Nafta Katowice Pol.* **37** (1981) 280; *Chem. Abstr.* **96** (1982) 125818m.
152. Monsanto, EP-A 165 227, 1985 (B. G. Beestman).
153. Cassella, DE-O 3 333 654, 1983 (U. Greiner, K. Albert); *Chem. Abstr.* **102** (1985) 221898a.
154. NHK Spring, DE-O 2 807 861, 1977 (T. Iwashita); *Chem. Abstr.* **89** (1978) 216338b.
155. Davy Bamag, DE 2 748 652, 1979 (H. Kreuzer).
156. Armour Industrial Chemical, US 3 533 470, 1970 (B. E. Marsh, B. Mosier).
157. Furukawa Electric, JP Kokai Tokyo Koho 78 99 299, 1977 (K. Yamazaki).
158. Bayer, DE-O 4 217 024, 1992; *Chem. Abstr.* **121** (1994) 11220q.

Design of Experiments

SERGIO SORAVIA, Process Technology, Degussa AG, Hanau, Germany (Chap. 1, 2, 3 and 8)
ANDREAS ORTH, University of Applied Sciences, Frankfurt am Main, Germany (Chap. 4, 5, 6, 7 and 8)

1. Introduction . . . 1
1.1. General Remarks . . . 1
1.2. Application in Industry . . . 2
1.3. Historical Sidelights . . . 3
1.4. Aim and Scope . . . 3
2. Procedure for Conducting Experimental Investigations: Basic Principles . . . 3
2.1. System Analysis and Clear Definition of Objectives . . . 4
2.2. Response Variables and Experimental Factors . . . 4
2.3. Replication, Blocking, and Randomization . . . 5
2.4. Interactions . . . 6
2.5. Different Experimental Strategies . . . 6
2.6. Drawback of the One-Factor-at-a-Time Method . . . 7
3. Factorial Designs . . . 8
3.1. Basic Concepts . . . 8
3.2. The 2^2 Factorial Design . . . 9
3.3. The 2^3 Factorial Design . . . 12
3.4. Fractional Factorial Designs . . . 15
4. Response Surface Designs . . . 18
4.1. The Idea of Using Basic Empirical Models . . . 19
4.2. The Class of Models Used in DoE . . . 19
4.3. Standard DoE Models and Corresponding Designs . . . 20
4.4. Using Regression Analysis to Fit Models to Experimental Data . . . 22
5. Methods for Assessing, Improving, and Visualizing Models . . . 22
5.1. R^2 Regression Measure and Q^2 Prediction Measure . . . 23
5.2. ANOVA (Analysis of Variance) and Lack-of-Fit Test . . . 24
5.3. Analysis of Observations and Residuals . . . 26
5.4. Heuristics for Improving Model Performance . . . 26
5.5. Graphical Visualization of Response Surfaces . . . 27
6. Optimization Methods . . . 28
6.1. Basic EVOP Approach Using Factorial Designs . . . 28
6.2. Model-Based Approach . . . 29
6.3. Multi-Response Optimization with Desirability Functions . . . 29
6.4. Validation of Predicted Optima . . . 30
7. Designs for Special Purposes . . . 31
7.1. Mixture Designs . . . 31
7.2. Designs for Categorical Factors . . . 33
7.3. Optimal Designs . . . 35
7.4. Robust Design as a Tool for Quality Engineering . . . 36
8. Software . . . 37
9. References . . . 37

1. Introduction

1.1. General Remarks

Research and development in the academic or industrial context makes extensive use of experimentation to gain a better understanding of a process or system under study. The methodology of *Design of Experiments (DoE)* provides proven strategies and methods of experimental design for performing and analyzing test series in a systematic and efficient way. All experimental parameters are varied in an intelligent and balanced fashion so that a maximum of information is gained from the analysis of the experimental results. In most cases, the time and money spent on the experimental investigation will be greatly reduced. In all cases, an optimal ratio between the number of experimental trials and the information content of the results will be achieved.

DoE is a powerful target-oriented tool. If it is properly employed, creative minds with a scientific and technical background will best deploy their resources to reach a well-defined

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.e08_e01.pub2

goal of their studies. In contrast to what researchers sometimes fear, experimenters will not be hampered in their creativity, but will be empowered for structuring their innovative ideas. Of course, adopting DoE requires discipline from the user, and it has proved very helpful to take the initial steps together with an expert with experience in the field. The rewards of this systematic approach are useful, reliable, and well-documented results in a clear time and cost frame. A comprehensible presentation and documentation of experimental investigations is gratefully acknowledged by colleagues or successors in research and development teams. The application of DoE is particularly essential and indispensable when processes involving many factors or parameters are the subject of empirical investigations of cause – effect relationships.

DoE is a scientific approach to experimentation which incorporates statistical principles. This ensures an objective investigation, so that valid and convincing conclusions can be drawn from an experimental study. In particular, an honest approach to dealing with process and measurement errors is encouraged, since experiments that are repeated under identical conditions will seldom lead to the same results. This may be caused by the measuring equipment, the experimenter, changes in ambient conditions, or the natural variability of the object under study. Note that this inherent experimental error, in general, comprises more than the bare repeatability and reproducibility of a measurement system. DoE provides basic principles to distinguish between experimental error and a real effect caused by consciously changing experimental conditions. This prevents experimenters from drawing erroneous conclusions and, as a consequence, from making wrong decisions.

1.2. Application in Industry

In industry, increasingly harsh market conditions force companies to make every effort to reach and maintain their competitive edge. This applies, in particular, in view of the following goals:

- Quality of products and services in conformance to market requirements
- Low costs to ensure adequate profits
- Short development periods for new or improved products and production processes (time to market)

Quality engineering techniques are powerful elements of modern quality management systems and make it possible to reach these goals. One important challenge in this context is not to ensure quality downstream at the end of the production line, but to ensure product quality by a stable and capable production process which is under control. By this means, ongoing tests and checks to prove that the product conforms to specification requirements are avoided. This can be realized by knowing the important and critical parameters or factors governing the system and through the implementation of intelligent process management strategies.

A methodical approach, sometimes referred to as off-line quality engineering [1, 2], focuses even further upstream. By considering quality-relevant aspects in the early stages of product and process development, quality is ensured preventively in terms of fault prevention [3, 4]. Naturally, there is considerable cost-saving potential in the early stages of product and process design, where manufacturing costs are fixed to a large extent. The losses incurred for each design modification of a product or process gradually increase with time. In addition, design errors with the most serious consequences are known to be committed in these early stages.

As an outstanding quality engineering tool, DoE occupies a key position in this context. The emphasis is on engineering quality into the products and processes. At the same time, DoE opens up great economic potential during the entire development and improvement period. It is well-known that the implementation and use of corresponding methods increases competitiveness [5 – 9]. DoE is applied successfully in all high-technology branches of industry. In the process industry, it makes essential contributions to optimizing a large variety of procedures and covers the entire lifecycle of a product, starting from product design in chemical research (e.g., screening of raw materials, finding the best mixture or formulation, optimizing a chemical reaction), via process development in process engineering (e.g., test novel technological solutions, determine best operating conditions, optimize the performance of processes), up to production

(e.g., start up a plant smoothly, find an operating window which meets customer requirements at low cost, high capacity, and under stable conditions) and application technology (give competent advice concerning the application of the product, customize the properties of products to specific customer needs). In particular, technologies like high-throughput screening or combinatorial synthesis with automated workstations require DoE to employ resources reasonably.

1.3. Historical Sidelights

The foundations of modern statistical experimental design methods were laid in the 1920s by R. A. Fisher [10] in Great Britain and were first used in areas such as agricultural science, biology, and medicine. By the 1950s and 1960s, some of these methods had already spread into chemical research and process development where they were successfully employed [11 – 13]. During this period and later, G. E. P. Box et al. made essential contributions to the advancement and application of this methodology [5, 14 – 18].

In 1960, J. Kiefer and J. Wolfowitz initiated a profound research of the mathematical theory behind optimal designs, and in the early 1970s the first efficient algorithms for so-called D-optimal designs were developed.

Around the same time, G. Taguchi integrated DoE methods into the product and process development of numerous Japanese companies [19]. One of his key ideas is the concept of robust design [1, 2, 4]. It involves designing products and processes with suitable parameters and parameter settings so that their functionality remains as insensitive as possible to unavoidable disturbing influences. Taguchi's ideas were discussed fruitfully in the United States during the 1980s [1, 5, 20, 21] and caused an increasing interest in this subject in Western industries. At the same time, another set of tools, which in general is not suitable for chemical or chemical engineering applications, became popular under the name of D. Shainin [22, 23].

The 1980s also saw the advent and spread of software for DoE. Various powerful software tools have been commercially available for several years now. They essentially support the generation, analysis, and documentation of experimental designs. Experience has shown that such software can be used by experimenters once the basic principles of the methods applied have been learned. A list of software tools is given in Chapter 8.

Despite these stimulating developments, the majority of scientists and engineers in industry and academic research have still not yet used DoE.

1.4. Aim and Scope

The target group of this article consists of scientists and engineers in industry or academia. The intention is

- To give an insight into the basic principles and strategies of experimental design
- To convey an understanding of the important design and analysis methods
- To give an overview of optimization techniques and some advanced methods
- To demonstrate basically the power of proven methods of DoE and to encourage their use in practice

References for further reading and detailed study are given on all subjects. In particular, [16, 24], and [25] may be regarded as standard monographs on DoE in their respective languages.

2. Procedure for Conducting Experimental Investigations: Basic Principles

DoE should become an integral part of the regular working tools of all experimenters. The methods provided by this approach may be used in experimental investigations on a smaller scale as well as in large projects which, depending on their importance and scope, may involve putting together an interdisciplinary team of appropriate size and composition. It proved to be advantageous to also include staff who take care of the equipment or facility on-site. They often bring in aspects and experiences with which the decision-makers involved are largely unfamiliar. Moreover, involving, for instance, laboratory staff in the planning phase of experiments has a positive effect on their motivation. A DoE

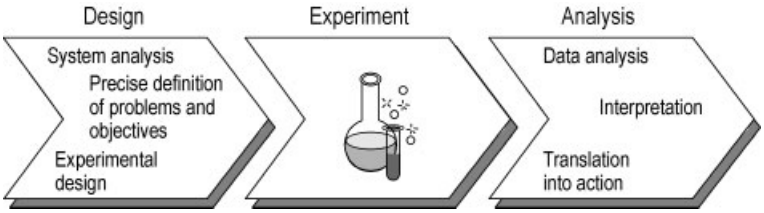

Figure 1. Three essential phases characterize the basic structure of a DoE project

project essentially subdivides into three characteristic phases: the design or planning phase, the actual experimental or realization phase, and the analysis or evaluation phase (Fig. 1). It should be stressed that the design phase is decisive because it is this phase that determines the level of information attainable by the analysis of the experimental results. For more information on this subject and on basic principles of DoE, see [26 – 28].

2.1. System Analysis and Clear Definition of Objectives

A system analysis involves collecting all existing information about the system to be examined and describing the current situation. A precise formulation of the problem and a clear definition of the objectives are very important prerequisites for a successful procedure and for attaining useful results from an experimental investigation. Experiments should never be conducted for their own sake but are to provide objective and reliable information, particularly, as a sound basis for decisions to be made. A clear statement of objectives is crucial, since the experimental strategy and hence the design of the experiments is essentially influenced by the goals to be reached (see Section 2.5). This sounds trivial but is frequently not handled carefully enough in practice. The actual planning or even the performance of experimental trials should not start until all of the aforementioned points have been settled satisfactorily.

2.2. Response Variables and Experimental Factors

Each experiment can be regarded as an inquiry addressed to a process or system (see Fig. 2). Naturally, it must be possible to record its answer or result in terms of measurable or quantifiable response variables (dependent variables). Response variables must be selected such that the characterization of the interesting properties of the system is as complete and simple as possible. Let us consider a batch reaction, for instance. In this case it is certainly not only the conversion of starting material that is of interest as a test result; other potential response variables could be the concentration of undesired byproducts or the duration of the reaction (capacity). For some system characteristics, such as foaming during stirring operations or the visual impression of a pigment, there are often no measurable quantities. In these cases, it may be helpful to assess the results by a subjective rating (e.g., 0: no foam or very beautiful pigment, 1: some foam or beautiful pigment, up to about 7: very much foam or very ugly pigment). When the response variables are selected, it is important to ask questions about the reliability of the corresponding values: How much do the results vary if the experimental runs are conducted under conditions that are as identical as possible? Ideally, a detailed analysis of the measurement system is available which, in particular, provides information on repeatability and reproducibility.

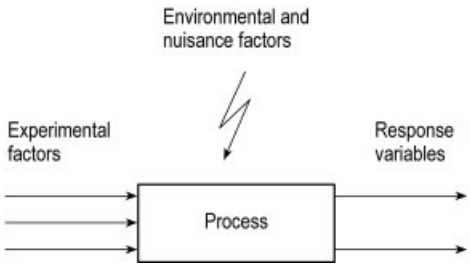

Figure 2. Input – output model of a process or system

The results of experimental runs, i.e., the values of response variables, are affected by various factors. In practically all applications there are disturbing environmental factors, which may cause undesired variations in the response variables. Some of them are hard to control or uncontrollable, others simply may not even be known. Examples of such variables are different batches of a starting material, various items of equipment with the same functionality, a change of experimenters, atmospheric conditions, and — last but not least — time-related trends such as warming-up of a machine, fouling of a heat exchanger, clogging of a filter, or drifting of a measuring instrument.

However, besides collecting and discussing the uncontrollable and disturbing variables, it is essential to collect and weigh up those factors or parameters that can be controlled or adjusted (independent variables), such as temperature, pressure, or the amount of catalyst used. The decision as to which of these experimental factors are to be kept constant during a test series and which are to be purposefully and systematically varied must also be carefully weighed up. The entire scientific and technical system know-how available by then from literature and experience, as well as intuition, must decisively influence not only the choice of the factors to be varied (one should focus on the important ones here according to the latest knowledge) but also the determination of the experimental region, i.e., of the specific range over which each factor will be varied (e.g., temperature between 120 and 180 °C and pressure between 1200 and 1800 mbar). A good experimental design will efficiently cover this experimental domain such that the questions related to the objectives may be answered when the experimental results are analyzed.

2.3. Replication, Blocking, and Randomization

To take the effects of disturbing environmental variables into account and decrease their impact to a large extent, the principles of *replication*, *blocking*, and *randomization* are employed. Moreover, by considering these principles, the risk of misleading interpretation of the results is minimized.

Replicates serve to make results more reliable and to obtain information about their variability. A genuine replicate should consider all possible environmental influences leading to variations within the response variables, i.e., the whole experiment with its corresponding factor-level combination should be repeated from the beginning and with some time delay in between. The results of data analyses must always be seen in the light of this inherent experimental error. Hence, replication does not mean, for instance, analyzing a sample of a product several times. This variability of a response variable is solely a measure of the precision of the analytical test procedure (laboratory assistant, measuring instrument).

It is of crucial importance that environmental factors and experimental factors of interest do not vary together such that changes in a response variable cannot be unambiguously attributed to the factors varied. For example, if two catalyst types were to be compared at various reaction temperatures and if two differently sized reactors were available for this purpose, it would be unwise to conduct all experiments with one catalyst in the smaller reactor and all experiments with the other catalyst in the larger reactor. If, in this case, the results differed from each other, it would be impossible to decide whether the catalyst type or the reactor type or both caused the deviations. The objective of blocking is to predetermine relatively similar blocks — in this case, the two reactors — in which test conditions are more homogeneous and which allow a more detailed study of the experimental factors of interest. Regarding the selection of the catalyst type and reaction temperature, the experiments to be conducted in each of the two reactors must be similar in the sense that variations in the values of a response variable can be interpreted correctly.

It is not always possible, however, to clearly identify unwanted influences and to take them into account, as is the case when blocking is used. Yet these side effects can be counterbalanced by a general use of randomization. Here, in contrast to systematically determining *which* experiments are to be conducted, *the order* of the experiments is randomized. In particular, false assignments of time-related trends are avoided. Let us consider a rectifying column, for instance, in which the effects of operating pressure, reflux

ratio, and reboiler duty on the purity of the top product are to be examined. Let us assume that the unit is started up in the morning and that the whole test series could be realized within one day. Now, if one conducted all experiments involving a low reflux ratio before those involving a high reflux ratio, the effect of the reflux ratio could be falsified more or less by the unit's warming up, depending on how strong this influence is (poor design in Fig. 3). Such an uncontrolled mixing of effects is prevented by choosing the order of the experimental runs at random (good design in Fig. 3).

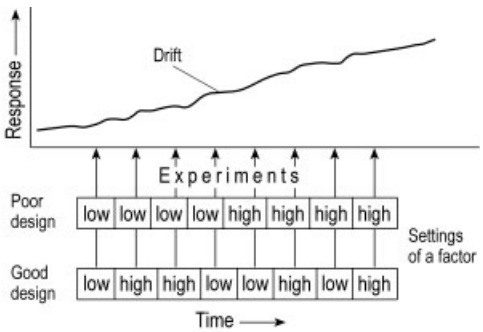

Figure 3. Time-related effects caused, e.g., by instrument drifts may falsify the analysis of the results when factor settings are not changed randomly (poor design). If environmental conditions vary during the course of an experiment, their effect will be damped or averaged out by randomizing the sequence of the experiments (good design)

The decisive reason for employing the principles of replication, blocking, and randomization is therefore to prevent the analysis of systematically varied experimental factors from being unnecessarily contaminated by the influences of unwanted and often hidden factors. While blocking and randomization basically do not involve additional experimental runs, each replicate is a completely new realization of a combination of factor settings. An appropriate relation between the number of experimental runs and the reliability of its results must be established here on an individual basis.

2.4. Interactions

To avoid an overly limited view on the behavior of systems, it is of great importance to know about the joint effects of experimental factors, that is, their *interactions*. Two variables are said to interact if, by changing one, the extent of impact on a third, namely, a response variable, depends on the setting of the other variable. In other words, interaction between two experimental factors measures how much the effect of a factor variation on a response variable depends on the level of the other factor. Interactions are often not heeded in practice, or they are studied at the price of spending large amounts of time and money on the associated experimental investigation. In addition, what interaction actually means is often not clearly understood. In particular, interaction is not to be confused with correlation. Two variables are said to be correlated if an increase of one variable tends to be associated with an increase or decrease of the other. Especially factorial experimental designs (see Chap. 3) allow, among other things, a quantitative determination of interactions between varied experimental factors.

2.5. Different Experimental Strategies

When processes are to be improved or novel technical solutions are to be tested, but also when plants are started up, several factors are often varied, in the hope of meeting with short-term success, by using an unsystematic iterative trial-and-error approach until satisfactory results are eventually produced. The expenditure involved quickly takes on unforeseeable proportions without affording important insights into the cause-and-effect relationships of the system. The result of this procedure is that, in the end, a comprehensible documentation is not available, and objective, reliable reasons for process operations or factor settings are missing. Furthermore, very little is known in most cases about the impact of factor variations. Experimenting in this way might be acceptable for orientation purposes in a kind of pre-experimental phase. However, one should switch to a judicious program as soon as possible.

To study causal relationships systematically, the experimental factors or parameters are usually varied separately and successively, and the values of a response variable (product, process, or quality characteristic), such as the yield of a chemical product, are shown in a diagram (see Fig. 4). This one-factor-at-a-time method (see

Section 2.6), however, provides only few insights into the subject under study because the effect of a particular factor is only known at a single factor-level combination of the other factors. The response variable may have quite another shape if the levels of the remaining factors are set differently. If the experimental factors in their effect on a response variable do not act additively according to the superposition principle, i.e., if the factors influence each other in their effect on the response variable by existing interactions, a misinterpretation of the results is easily possible, particularly when optimum conditions are to be attained.

When statistical experimental design methods are used, all considered factors are varied in a systematic and balanced way so that a maximum of information is gained from the analysis of the corresponding experiments. This may comprise the statistically sound quantitative determination of the effects of factor variations on one or several response variables (see Chap. 3) or a systematic optimization of factor settings (see Chap. 6). Depending on the experimenter's intention, the following questions can be answered:

- What are the most important factors of the system under investigation?
- To what extent and in which direction does a response variable of interest change when an experimental factor is varied?
- To what extent is the size and direction of the effect of a factor variation dependent on the settings of other experimental factors (interactions)?
- With which factor settings does one obtain a desired state of a response variable (maximum, minimum, nominal value)?
- How can this state be made insensitive to disturbing environmental factors or how can an undesired variability of a response variable be reduced (robust design)?

The question of which experimental strategy should be chosen from a comprehensive range of methods will be governed by the objectives to be achieved in each individual case, taking, e.g., system-inherent, financial, and time-related boundary conditions into account. Every project has its peculiarities. Carefully planned experiments cover the experimental region to be investigated as evenly as possible, while ensuring to the largest possible extent that changing values in a response variable can be attributed unambiguously to the right causes. Information of crucial importance is frequently obtained by a simple graphical analysis of the data without having to employ sophisticated statistical analysis methods, such as variance analysis or regression analysis. On the other hand, the best statistical analysis is not capable of retrieving useful information from a badly designed series of experiments. It is therefore decisive to consider basic DoE principles right from the beginning, above all, however, before conducting any experiments.

2.6. Drawback of the One-Factor-at-a-Time Method

A crystallization process is used in the following to illustrate the deficiency of the frequently used one-factor-at-a-time method. Factors influencing this system are, for instance, crystallization conditions such as geometry of the crystallizer, type and speed of the agitator, temperature, residence time, and concentrations of additives like crystallization and filter aids, as well as of two presumed additives A and B. Possible response variables may be bulk density, abrasion, hardness, and pourability of the crystallization product. Let us assume the simple case that the effects of the two experimental factors — additive A and additive B — on the material's bulk density are to be systematically examined with the aim of obtaining a maximum bulk density. As mentioned before, the experimental factors are usually examined and/or optimized separately and successively. In the example considered here, one would therefore begin by keeping factor B constant and varying A over a certain range until A has been optimally adjusted in terms of a maximum bulk density and enter the result in a diagram (see Fig. 4). The optimal value for A would then be selected and kept constant. The same procedure would then be employed for B. The result of this is a presumably optimal setting for A and B, and hasty experimenters would jump to the conclusion that, in this case, after varying A between 15 and 40 g/L and B between 5 and 17.5 g/L, the highest bulk density is obtained by setting A to 33 g/L and B to 8.5 g/L and that its value is approximately 825

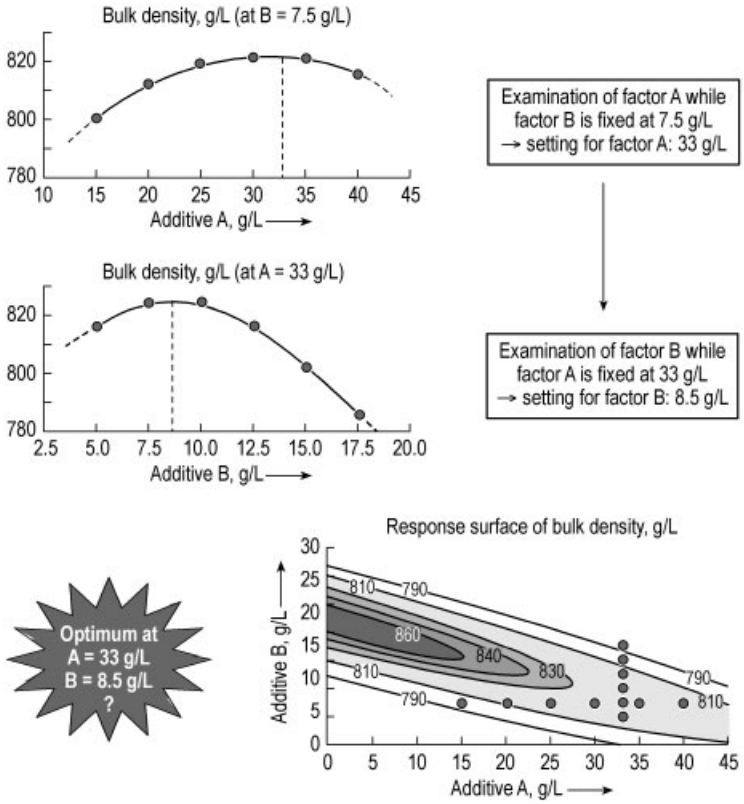

Figure 4. The one-factor-at-a-time method can lead to misinterpretations in systems that are subject to interactions

g/L. However, the response surface in Figure 4, which shows the complete relationship between both experimental factors and the response variable, reveals how misleading such a conclusion can be. This drastic misinterpretation is based on the, in this case, false assumption that the effect of varying one factor is independent of the settings of the other factor. For instance, by using the response surface, one can show that the bulk density values take on a decidedly different shape compared to the first diagram when varying A for B = 15 g/L. The following should be noted: If there are interactions between experimental factors, the one-factor-at-a-time method is an unsuitable tool for a systematic analysis of these factors, which holds true in particular when factor settings are to be optimized. If such interactions can definitely be ruled out, it might well be used. In chemistry, however, it is rather the rule that interactions occur.

3. Factorial Designs

3.1. Basic Concepts

The statistical experimental designs most frequently used in practice are the *two-level factorial designs*. These designs are called two-level because there are only two levels of settings, a lower (−) and an upper level (+), for each of the experimental factors. A full two-level factorial design specifies all combinations of the lower and upper levels of the factors as settings of the experimental runs (2^n design, where n denotes the number of factors). Their principle is illustrated by a simple example of a chemical reaction for which the influence of 2 (2^2 design) and 3 (2^3 design) experimental factors on the product yield is to be examined (Sections 3.2 and 3.3). For a growing number of factors, the number of runs of a full factorial design increases exponentially, and it provides much more information than is generally needed. Particularly for $n > 4$, the number of experimental settings can

be reduced by selecting a well-defined subgroup of all 2^n possible settings of the factors without losing important information. This leads to the fractional factorial designs (Section 3.4).

The restriction of initially using just two levels for each experimental factor often causes some uneasiness for experimenters using this method for the first time. But by using two-level factorial designs, a balanced coverage of the interesting experimental region is achieved very economically. Moreover — owing to the special combination of factor levels — it is also possible to gain deeper insights from the associated individual values of the response variables. A decisive advantage of the two-level factorial designs is that they allow the effects of factor variations to be systematically and reliably analyzed and quantified and that they provide information on how these effects depend on the settings of the other experimental factors. These insights are gained by calculating so-called *main effects* and *interaction effects*. In the calculation of these effects, all experimental results can be used and are included to form well-defined differences of corresponding averages (see Figs. 6 and 9), thereby increasing the degree of reliability. The essential results of this effect analysis can be visualized by simple diagrams (see, e.g., Fig. 7).

In a factorial design, not only continuous experimental factors, such as temperature, pressure, and concentration, which can be set to any intermediate value, but also discrete or categorical factors, such as equipment or solvent type, may be involved. If at least one categorical variable with more than two levels is involved or if curvatures in the response variables are expected and to be explored, factorial designs with more than two levels may be used, e.g., 3^n designs, in which all factors are studied at three levels each, or hybrid factorial designs with mixed factor levels like the 2×3^2 design, in which one factor is varied at two levels, and two factors at three levels [24]. However, especially in the case of continuous factors, other so-called response surface designs are more efficient (see Chap. 4). In the following, the expression "factorial designs" always refers to two-level factorial designs.

For the sake of simplicity, replicates are neglected in the following examples, and variability in the process and in measurement are assumed to be very small. Note, however, that being aware of the impact of experimental error on the reliability or significance of calculated effects is an essential principle of DoE and crucial to drawing valid conclusions. Variability within individual runs having the same settings of the experimental factors will propagate and cause variability in each calculated variable, e.g., main effect or interaction effect, deduced from these single results. Experimental designs, particularly the factorial designs, minimize error propagation.

Factorial designs are treated in most textbooks on DoE, e.g., [16, 24, 25, 29, 30].

3.2. The 2^2 Factorial Design

Let us suppose the influence of two factors — catalyst quantity A and temperature B — on yield y of a product in a stirred tank reactor is to be examined. Figure 5 shows the two levels of both factors involved, as well as a tabular and a graphical representation of the associated experimental 2^2 factorial design. The — in this case two — columns which contain the settings of the experimental factors form the so-called design matrix. The resultant values of the response variable y obtained for the settings A−B−, A+B−, A−B+, and A+B+ are referred to as $y_{\mathrm{A-B-}}$, $y_{\mathrm{A+B-}}$, $y_{\mathrm{A-B+}}$, and $y_{\mathrm{A+B+}}$ respectively. They are entered in the column of the response variable and in the corresponding positions of the graph.

Due to the special constellation of the experimental runs, it is possible to see how y changes when factor A is varied at the two levels of B and what happens when factor B is varied at the two levels of A. Figure 5 reveals that, at the lower temperature of 70 °C (B−), an increase of the catalyst quantity from 100 g (A−) to 150 g (A+) increases the response variable yield by 2 %, while at the higher temperature of 90 °C (B+), increasing the catalyst quantity enlarges the value of the response variable by 18 %, i.e.,

Effect of A for B−
$= \mathrm{E}_{\mathrm{B-}}(\mathrm{A})$
$= y_{\mathrm{A+B-}} - y_{\mathrm{A-B-}}$
$= +2\%$
Effect of A for B+
$= \mathrm{E}_{\mathrm{B+}}(\mathrm{A})$
$= y_{\mathrm{A+B+}} - y_{\mathrm{A-B+}}$
$= +18\%$

ID	Factor	Unit	Lower level (-)	Upper level (+)
A	Catalyst	g	100	150
B	Temperature	°C	70	90

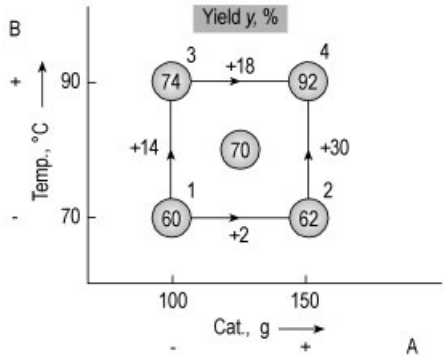

No.	A	B	y
1	-	-	60
2	+	-	62
3	-	+	74
4	+	+	92

Figure 5. Example of a 2^2 factorial design, including a graphical and a tabular representation of the experimental settings and results

Accordingly, the individual effects of a temperature increase can be determined for the different catalyst quantities:

Effect of B for A−

$$=\mathrm{E}_{\mathrm{A}-}(\mathrm{B})$$
$$=y_{\mathrm{A-B+}}-y_{\mathrm{A-B-}}$$
$$=+14\%$$

Effect of B for A+

$$=\mathrm{E}_{\mathrm{A}+}(\mathrm{B})$$
$$=y_{\mathrm{A+B+}}-y_{\mathrm{A+B-}}$$
$$=+30\%$$

As mentioned before, key results of a factorial design are obtained by the calculation of main effects and interaction effects (Fig. 6). The main effect of a factor is a measure of the extent to which a response variable changes on average when this factor is varied from its lower to its upper level. To calculate this main effect, all experimental results obtained at the upper (+) and lower (−) level of the factor are averaged, and the result at the lower setting is subsequently subtracted from that at the upper. This leads to the following equations:

Main effect of A

$$=\mathrm{ME}(\mathrm{A})$$
$$=\frac{y_{\mathrm{A+B-}}+y_{\mathrm{A+B+}}}{2}-\frac{y_{\mathrm{A-B-}}+y_{\mathrm{A-B+}}}{2}$$
$$=\frac{(y_{\mathrm{A+B-}}-y_{\mathrm{A-B-}})+(y_{\mathrm{A+B+}}-y_{\mathrm{A-B+}})}{2}$$
$$=\frac{\mathrm{E}_{\mathrm{B-}}(\mathrm{A})+\mathrm{E}_{\mathrm{B+}}(\mathrm{A})}{2}$$
$$=10\%$$

Main effect of B

$$=\mathrm{ME}(\mathrm{B})$$
$$=\frac{y_{\mathrm{A-B+}}+y_{\mathrm{A+B+}}}{2}-\frac{y_{\mathrm{A-B-}}+y_{\mathrm{A+B-}}}{2}$$
$$=\frac{(y_{\mathrm{A-B+}}-y_{\mathrm{A-B-}})+(y_{\mathrm{A+B+}}-y_{\mathrm{A+B-}})}{2}$$
$$=\frac{\mathrm{E}_{\mathrm{A-}}(\mathrm{B})+\mathrm{E}_{\mathrm{A+}}(\mathrm{B})}{2}$$
$$=22\%$$

The expressions preceding the numerical results in the calculations above show that the main effect of a factor can also be calculated from the mean of the individual effects involved. The two calculated results can now be read as follows: When the catalyst quantity A is increased from 100 to 150 g, the yield increases on average by 10 % from $\frac{60+74}{2}=67\%$ to $\frac{62+92}{2}=77\%$. The temperature B has a stronger impact on the response variable yield within the range of 70 – 90 °C. An increase in temperature leads to an average increase in yield of 22 %. The main-effect diagrams in Figure 7 illustrate these relations graphically.

The effect of a factor variation is strongly dependent on the respective setting of the other factor. This was already shown in Figure 5 and is further illustrated by the two interaction diagrams in Figure 7. The four lines in these diagrams correspond to the four edges of the response surface, also shown in Figure 7. This interaction is a typical case of a synergetic interaction. A simultaneous increase of A and B has

a clearly higher impact on the response variable than the additive superposition of the individual effects, $E_{B-}(A)$ and $E_{A-}(B)$, would lead one to expect. If the value of y_{A+B+} were not 92 but 76 %, there would be no interaction between A and B. The effect of a variation of one of the two factors on the response variable would be independent of the adjustment of the other factor. In this case, the two individual effects would be identical for each factor, i.e., $E_{B-}(A) = E_{B+}(A) = +2\,\%$ and $E_{A-}(B) = E_{A+}(B) = +14\,\%$. The corresponding lines in the interaction diagrams would then be parallel. It seems reasonable now to determine a quantitative measure for the interaction of two factors as the difference of these individual effects, i.e.,

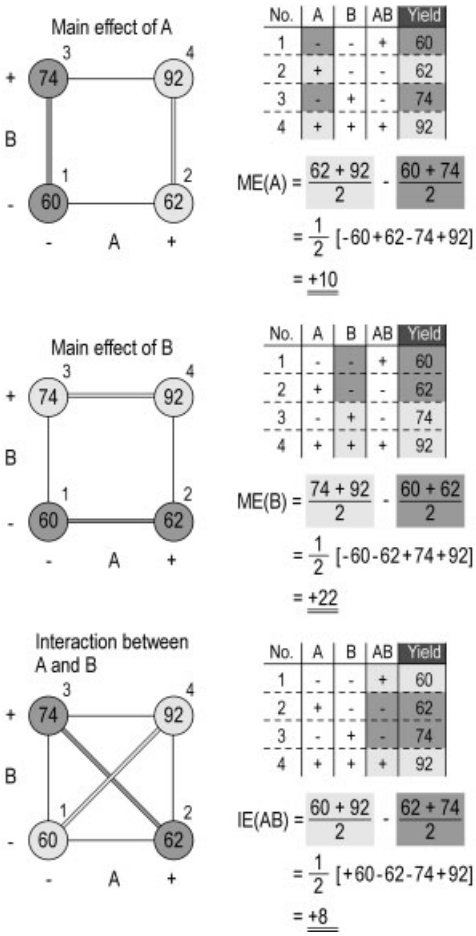

Figure 6. Calculation of the two main effects and the interaction effect in a 2^2 factorial design

Interaction effect between A and B

$$
\begin{aligned}
&= \mathrm{IE}\,(\mathrm{AB}) \\
&= \tfrac{1}{2}\,[\mathrm{E}_{B+}\,(\mathrm{A}) - \mathrm{E}_{B-}\,(\mathrm{A})] \\
&= \frac{(y_{A+B+} - y_{A-B+}) - (y_{A+B-} - y_{A-B-})}{2} \\
&= \frac{y_{A+B+} + y_{A-B-}}{2} - \frac{y_{A-B+} + y_{A+B-}}{2} \\
&= \frac{(y_{A+B+} - y_{A+B-}) - (y_{A-B+} - y_{A-B-})}{2} \\
&= \tfrac{1}{2}\,[\mathrm{E}_{A+}\,(\mathrm{B}) - \mathrm{E}_{A-}\,(\mathrm{B})] \\
&= \mathrm{IE}\,(\mathrm{BA}) \\
&= +8\%
\end{aligned}
$$

From these sequences of expressions it can be seen that the interaction between A and B could be equally defined through the difference of the individual effects of A or the individual effects of B, giving the same numerical result each time. Moreover, the expression in the middle shows that the interaction between A and B — as is the case for the main effects — is nothing but the difference of two averages (this is basically the reason why a factor of $\frac{1}{2}$ is introduced in the definition). This corresponds to the difference of the averages of the results located on the diagonals in Figure 6.

The calculation of the two main effects and of the interaction effect is also represented geometrically in Figure 6. The corresponding analysis table contains columns of signs, which allow calculation of these effects. Each effect may be computed as the sum of signed response values divided by half the number of experiments, where the signs are taken from the column of the desired effect. Note that the signs of the interaction column AB can be generated by a row-wise multiplication of the main-effect columns A and B. Today, however, the effects do not need to be computed like this anymore. Specific DoE software tools (see Chap. 8) use methods such as those described in Section 4.4 and yield numerical and graphical results more easily.

A final interpretation of the experiments could read as follows: The catalyst shows a not yet satisfactory activity at the lower temperature (70 °C). An increase of the catalyst quantity from 100 to 150 g gives a slight improvement but does not yet yield satisfactory values. The situation is different at the higher temperature (90 °C), where the catalyst clearly performs better. In addition, an increase in the catalyst quantity at this temperature has a strong impact on yield.

If all experimental factors are continuous, it will be possible and useful to perform experi-

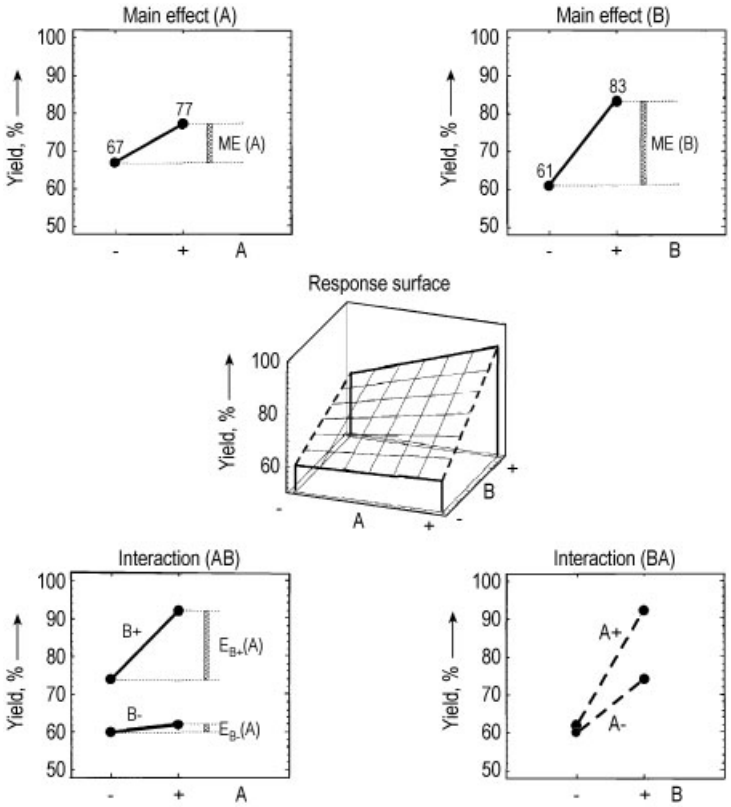

Figure 7. Diagrams of main effects, interactions, and response surfaces illustrate conspicuously important relations governing the system

mental runs at the center point. It is obtained by setting each factor to the midpoint of its factor range, in our example, to 125 g of catalyst and 80 °C. This isolated additional experimental point gives a rough impression of the behavior of response variables inside the experimental region of interest. If the result in the center point does not correspond to the mean of the results obtained in the corner points of the factorial design, then the response surface of the response variable will have a more or less pronounced curvature that depends on the magnitude of this deviation. The graph in Figure 5 shows a result of 70 % at the center point, which is slightly below the 72 % obtained by averaging the four results of the factorial design. Thus, the response surface must be imagined as slightly sagging in its middle. Of course, with this single additional experiment, it is impossible to determine which of the experimental factors is (are) ultimately responsible for the curvature. This question can only be settled by a response surface design (see Chap. 4).

The analysis shown for one response variable is performed for each response variable so that the effects of factor variations on every response variable are finally known.

3.3. The 2^3 Factorial Design

The concepts and notions introduced in the Section 3.2 can be generalized to three or more factors. Let us suppose that, in addition to the example in Section 3.2, not only the effects of temperature and catalyst quantity but also the impact of changing the agitator type on the yield are to be examined. In analogy to Figure 5, the experimental factors with their respective two settings as well as the 2^3 factorial design, which is obtained by realizing all possible combinations of factor settings, are represented in graphical and tabular form in Figure 8.

ID	Factor	Unit	Lower level (-)	Upper level (+)
A	Catalyst	g	100	150
B	Temperature	°C	70	90
C	Agitator type		current	new

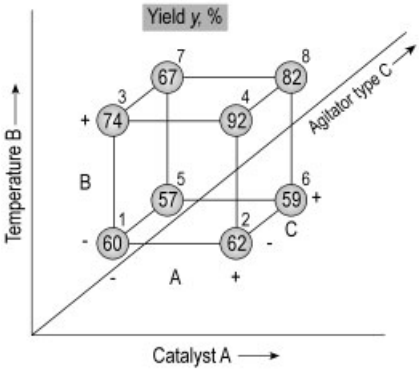

No.	A	B	C	y
1	-	-	-	60
2	+	-	-	62
3	-	+	-	74
4	+	+	-	92
5	-	-	+	57
6	+	-	+	59
7	-	+	+	67
8	+	+	+	82

Figure 8. Example of a 2^3 factorial design, including a graphical and a tabular representation of the experimental settings and results. The experimental settings for agitator type "current" correspond to the 2^2 factorial design in Figure 5

By comparing the values at the ends of the various edges of the cube in Figure 8, it is possible to perform a very elementary analysis. For each factor, the effect of its variation can be studied for the four different constellations of the other two factors. For example, the change from the current to the new agitator type does not lead to the presumed improvement in yield. This may be verified by looking at the four edges going from the front face to the back face of the cube. The deterioration is particularly severe at high temperature where the yield decreases from 74 to 67 % (for the lower level of catalyst) and from 92 to 82 % (for the upper level of catalyst).

More detailed information about the specific and joint effects of the factors is obtained by calculation of the main effects and interaction effects introduced in Section 3.2.

The main effect thereby indicates how much a response variable is affected on average by a variation of a factor and is measured as the difference in the average response for the two factor levels. Figure 9 illustrates this for the factors A (catalyst) and B (temperature). The main effect for C (agitator type) is obtained analogously by calculating the difference of the two averages at the back face and the front face of the cube.

The interaction effect of two factors, more precisely, the two-factor interaction was introduced in Section 3.2. Generally, there will be

$$\binom{n}{2} = \frac{n\cdot(n-1)}{2}$$

two-factor interactions, where n denotes the number of factors. For three factors there are

$$\binom{3}{2} = 3$$

two-factor interactions, namely, AB, AC, and BC. The calculation of a two-factor interaction in a 2^3 factorial design will be demonstrated for AB. This interaction is obtained by calculating the two-factor interaction $\mathrm{IE}_{\mathrm{C}-}(\mathrm{AB})$ at the lower level of factor C and the two-factor interaction $\mathrm{IE}_{\mathrm{C}+}(\mathrm{AB})$ at the upper level of factor C, and then by averaging these two values:

$$\begin{aligned}
&\text{Interaction effect between A and B}\\
&= \mathrm{IE}\,(\mathrm{AB})\\
&= \frac{\mathrm{IE}_{\mathrm{C}-}(\mathrm{AB}) + \mathrm{IE}_{\mathrm{C}+}(\mathrm{AB})}{2}\\
&= \frac{\left(\frac{60+92}{2} - \frac{62+74}{2}\right) + \left(\frac{57+82}{2} - \frac{59+67}{2}\right)}{2}\\
&= \frac{60+92+57+82}{4} - \frac{62+74+59+67}{4}\\
&= +7.25\%
\end{aligned}$$

A two-factor interaction in a factorial design with more than two factors is obtained by taking the average of all individual two-factor interactions at the different constellations of the other factors. The calculation of IE(AB) is also illustrated in Figure 9, where it is seen to be the difference of averages between results on two diagonal planes. Due to the inherent symmetry

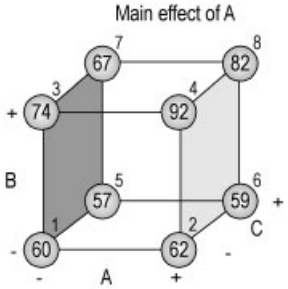

No.	A	B	C	AB	AC	BC	ABC	Yield
1	-	-	-	+	+	+	-	60
2	+	-	-	-	-	+	+	62
3	-	+	-	-	+	-	+	74
4	+	+	-	+	-	-	-	92
5	-	-	+	+	-	-	+	57
6	+	-	+	-	+	-	-	59
7	-	+	+	-	-	+	-	67
8	+	+	+	+	+	+	+	82

$$ME(A) = \frac{62+92+59+82}{4} - \frac{60+74+57+67}{4}$$

$$= \frac{1}{4}\,[-60+62-74+92-57+59-67+82]$$

$$= \underline{\underline{+9.25}}$$

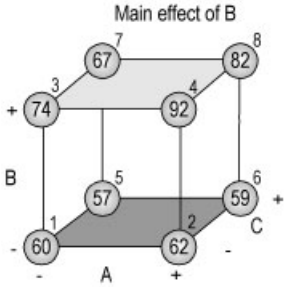

No.	A	B	C	AB	AC	BC	ABC	Yield
1	-	-	-	+	+	+	-	60
2	+	-	-	-	-	+	+	62
3	-	+	-	-	+	-	+	74
4	+	+	-	+	-	-	-	92
5	-	-	+	+	-	-	+	57
6	+	-	+	-	+	-	-	59
7	-	+	+	-	-	+	-	67
8	+	+	+	+	+	+	+	82

$$ME(B) = \frac{74+92+67+82}{4} - \frac{60+62+57+59}{4}$$

$$= \frac{1}{4}\,[-60-62+74+92-57-59+67+82]$$

$$= \underline{\underline{+19.25}}$$

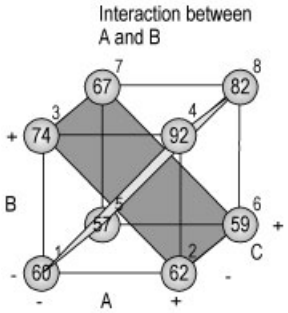

No.	A	B	C	AB	AC	BC	ABC	Yield
1	-	-	-	+	+	+	-	60
2	+	-	-	-	-	+	+	62
3	-	+	-	-	+	-	+	74
4	+	+	-	+	-	-	-	92
5	-	-	+	+	-	-	+	57
6	+	-	+	-	+	-	-	59
7	-	+	+	-	-	+	-	67
8	+	+	+	+	+	+	+	82

$$IE(AB) = \frac{60+92+57+82}{4} - \frac{62+74+59+67}{4}$$

$$= \frac{1}{4}\,[+60-62-74+92+57-59-67+82]$$

$$= \underline{\underline{+7.25}}$$

Figure 9. Calculation of main effects and interaction effects in a 2^3 factorial design shown for the main effects of A and B and their two-factor interaction AB

of the design, the calculation of the other two-factor interactions, IE(AC) and IE(BC), can be performed in a similar way. This leads to the respective columns of signs within the analysis table in Figure 9 and to the corresponding diagonal planes within the cube.

Now, if the interaction between A and B at the lower level of C differs from their interaction at the upper level of C, there will be a three-factor interaction, which is defined by the difference of these two individual two-factor interactions:

$$\begin{aligned}
&\mathrm{IE}\,(\mathrm{ABC})\\
&=\tfrac{1}{2}\ [\mathrm{IE}_{\mathrm{C}+}\,(\mathrm{AB})-\mathrm{IE}_{\mathrm{C}-}\,(\mathrm{AB})]\\
&=\tfrac{1}{2}\ \left[\left(\tfrac{57+82}{2}-\tfrac{59+67}{2}\right)-\left(\tfrac{60+92}{2}-\tfrac{62+74}{2}\right)\right]\\
&=\tfrac{62+74+57+82}{4}-\tfrac{60+92+59+67}{4}\\
&=-0.75\%
\end{aligned}$$

Note that this result can also be obtained by using the column of signs corresponding to the three-factor interaction ABC in Figure 9. As in the case of two-factor interactions, the signs within this column are obtained by a row-wise multiplication of the signs of the main effect columns A, B, and C. Obviously, the three-factor interaction is also the difference of two averages (as in the case of the two-factor interaction presented in Section 3.2, this is the reason why the factor of $\frac{1}{2}$ is introduced in the definition). They are obtained by averaging the results located at the vertices of the respective two tetrahedra which make up the cube.

The numerical results of all effects obtainable from the 2^3 factorial design are summarized in the following:

ME(A)	=	9.25 %
ME(B)	=	19.25 %
ME(C)	=	− 5.75 %
IE(AB)	=	7.25 %
IE(AC)	=	− 0.75 %
IE(BC)	=	− 2.75 %
IE(ABC)	=	− 0.75 %

Note again that numerical results like these and graphical results like those of Figure 7 are easily obtained by using specific DoE software (see Chap. 8). Diagrams of main effects and interaction effects may also be generated for factorial designs with more than two factors.

Even though three-factor interactions may really exist in some cases, it is seldom that they play an essential role. So, if their absolute value is clearly higher than most of the main effects and two-factor interactions, it seems reasonable to conclude that the experimental error is of a magnitude that does not allow the reliable estimation of most of the effects (perhaps because they are very small) and/or that at least one response value has been corrupted by a gross systematic error.

If, on the whole, systematic errors can be excluded, it is a legitimate practice to neglect higher-order interactions, such as three- and four-factor interactions, because main effects tend to be larger than two-factor interactions, which in turn tend to be larger than three-factor interactions, and so on. Moreover, if no information about the magnitude of the experimental error is available, it will be possible to obtain a rough estimate of this error by using higher-order interactions, like five-, four-, and even three-factor interactions.

3.4. Fractional Factorial Designs

For a growing number n of experimental factors, the number of experimental settings 2^n increases exponentially in a full factorial design. Simultaneously, the proportion of higher-order interactions increases rapidly. For instance, if $n = 5$, there are 5 main effects, 10 two-factor interactions, and 16 interactions of higher order. Obviously, if n is not small, there is some redundancy in a full factorial design, since higher-order interactions are not likely to have appreciable magnitudes. At this point, the following questions arise:

- Is it possible to reduce the amount of experimental effort in a sophisticated way so that the most important information can still be obtained by analysis of the data?
- Is it possible to study more experimental factors instead of higher-order interactions with the same number of experimental settings?

The answer in both cases is "yes". It leads to the 2^{n-k} fractional factorial designs, where 2^{n-k} denotes the number of experimental settings, n the number of experimental factors, and k the number of times by which the number of settings has been halved compared to the corresponding complete 2^n design with the same number of experimental factors ($1/2^k \cdot 2^n = 2^{n-k}$). The application of fractional factorial designs yields a reduction in experimental effort, which is adapted to the complexity of the system under investigation and to the information required.

Once the experimental runs have been performed, fractional factorial designs are analyzed like full factorial designs, except that fewer values are available. As introduced in Sections 3.1, 3.2, and 3.3, effects are obtained by calculating corresponding differences of averages. However, by doing so, one will discover that effects are confounded. *Confounding* is defined as a situation where an effect cannot unambiguously be attributed to a single main effect or interaction.

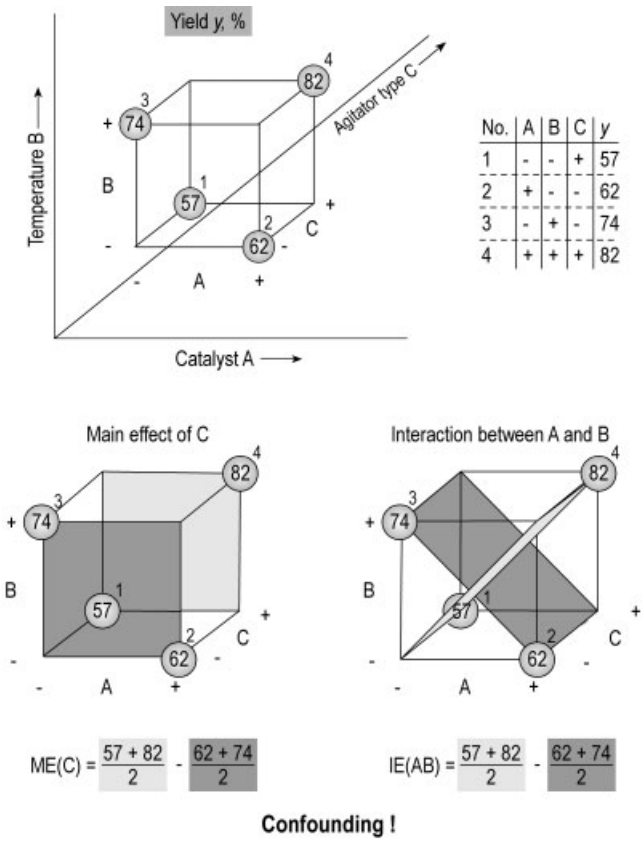

No.	A	B	C	y
1	-	-	+	57
2	+	-	-	62
3	-	+	-	74
4	+	+	+	82

Figure 10. Example of a 2^{3-1} fractional factorial design. In the top section the experimental settings and results are shown in geometrical and tabular form. In the bottom section the problem of confounding is illustrated: Calculating ME(C) and IE(AB) leads to the same expression, which is actually the sum of both. It is not possible to determine whether the calculated value is caused by the main effect of C, by the interaction effect of A and B, or by both

Let us consider the 2^{3-1} design, in which the effects of three factors are studied with four experimental settings (see Figure 10). This is an example of a half-fraction factorial design (see Figure 11). By calculating the main effect of C [ME(C) on the left-hand side of Fig. 10] the outcome is not the difference between the averages of four (as in the case of the 2^3 design) but the averages of only two results which are calculated. A similar situation occurs when calculating the interaction between A and B [IE(AB) on the right-hand side of Fig. 10]. Moreover, calculating these effects leads to the same expression. It is actually the sum ME(C)+IE(AB) of both effects. It can also be verified that ME(A) is confounded with IE(BC), and ME(B) with IE(AC). This is an example of a so-called resolution III design.

A slightly different situation occurs when considering the 2^{4-1} design (see Fig. 11). Here the main effects are confounded with three-factor interactions, e.g. ME(A) with IE(BCD), and the two-factor interactions are confounded with each other, e.g., IE(AC) with IE(BD). This is an example of a resolution IV design. The 2^{5-1} design also shown in Figure 11 is an example of a very efficient resolution V design.

The *resolution* of a fractional factorial design largely characterizes the information content obtainable by analyzing the results of a fractional factorial design:

- A fractional factorial design of resolution III does not confound main effects with each other but does confound main effects with two-factor interactions.
- A fractional factorial design of resolution IV does not confound main effects with two-

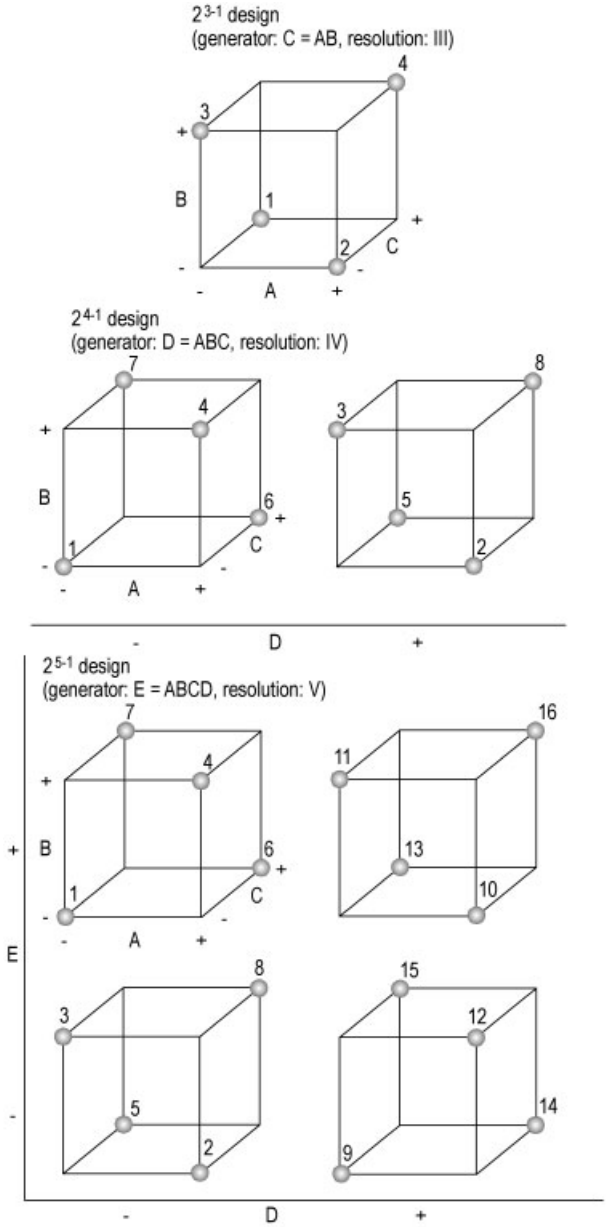

No.	A	B	C
1	-	-	+
2	+	-	-
3	-	+	-
4	+	+	+

No.	A	B	C	D
1	-	-	-	-
2	+	-	-	+
3	-	+	-	+
4	+	+	-	-
5	-	-	+	+
6	+	-	+	-
7	-	+	+	-
8	+	+	+	+

No.	A	B	C	D	E
1	-	-	-	-	+
2	+	-	-	-	-
3	-	+	-	-	-
4	+	+	-	-	+
5	-	-	+	-	-
6	+	-	+	-	+
7	-	+	+	-	+
8	+	+	+	-	-
9	-	-	-	+	-
10	+	-	-	+	+
11	-	+	-	+	+
12	+	+	-	+	-
13	-	-	+	+	+
14	+	-	+	+	-
15	-	+	+	+	-
16	+	+	+	+	+

Figure 11. Geometric representation of the most important half-fraction factorial designs and their respective tabular representation, i.e., their design matrices

factor interactions but does confound two-factor interactions with other two-factor interactions.

– A fractional factorial design of resolution V does not confound main effects and two-factor interactions with each other but does confound two-factor interactions with three-factor interactions.

The resolution of the most important fractional factorial designs can be seen in Table 1.

Designs of resolution III are used in the early stages of experimental investigations to gain a first insight into the possibilities and behaviors

Table 1. The 2^{n-k} (fractional) factorial designs with a maximum of 32 experimental settings and their resolution. 2^{n-k} denotes the number of experimental settings, n the number of experimental factors, and $1/2^k$ the factor by which the number of settings has been reduced compared to the corresponding full factorial design with the same number of experimental factors.

Resolution	Number of experimental settings			
	4	8	16	32
III	2^{3-1}	$2^{5-2} - 2^{7-4}$	$2^{9-5} - 2^{15-11}$	$2^{17-12} - 2^{31-26}$
IV		2^{4-1}	$2^{6-2} - 2^{8-4}$	$2^{7-2} - 2^{16-11}$
V			2^{5-1}	
VI				2^{6-1}
Full	2^2	2^3	2^4	2^5

of systems. Particularly beneficial are *saturated designs* for $2^n - 1$ factors in which all degrees of freedom of a 2^n design are exploited. The above-mentioned 2^{3-1} design is an example of such a design in which three factors are studied with four experimental settings. It is "generated" by replacing all columns of signs in the analysis table of Figure 6 by experimental factors, i.e., the settings of factor C at the top of Figure 11 are determined by the column of the two-factor interaction AB in the analysis table of Figure 6. A further important example of a saturated design is the 2^{7-4} design, which allows seven factors to be studied with eight experimental settings. It is generated by replacing AB by D, AC by E, BC by F, and ABC by G in the heading of the columns of signs used for the calculation of effects in Figure 9. With resolution III designs special care must be taken when interpreting the analysis results. Calculated main effects may also be two-factor interactions of other factors.

Resolution IV designs are employed to gain unambiguous information about the individual impact of the experimental factors, while unambiguous information about their two-factor interactions is not yet required. Designs of resolution III and IV are mostly used to find out which factors play an important role. This technique of isolating the important factors is sometimes referred to as *screening*.

Using designs of resolution V or higher does not lead to any loss of decisive information, since the main effects and two-factor interactions are only confounded with higher-order interactions, which in most cases can be neglected. Particularly when $n > 4$, experimental settings can be halved to achieve designs of at least resolution V.

Fractional factorial designs support the iterative nature of experimentation. The experimental runs of two or more fractional factorial designs conducted sequentially may be combined to form a larger design with higher resolution. In this way it is possible to resolve ambiguities by the addition of further experimental runs.

Half-fraction designs possess an interesting and useful projection property: the omission of one arbitrary column in the designs always leads to a full factorial design with respect to the remaining columns or factors. So, if a factor proves to have no significant effect on a response variable, the remaining factors can be analyzed as in the full factorial design. For example, the omission of factor B in the 2^{4-1} design of Figure 11 leads to a 2^3 design for the factors A, C, and D. This may be verified by examining the three remaining columns in the table but also by pushing the upper faces of the two cubes into the lower faces.

4. Response Surface Designs

In some situations factorial designs in which all factors are varied by only using two settings are not adequate for describing the behavior of an experimental system, because a more detailed insight is needed to predict its responses or to find optimal factor settings. In this case, it is often necessary to extend factorial designs and to do additional experiments at other points in the experimental domain. To decide which points to use, *response surface designs* are used [14, 31]. These designs are based on mathematical models that describe how responses depend on experimental factors.

4.1. The Idea of Using Basic Empirical Models

A model is a way to describe a part of reality; ideally it is much simpler and more manageable than that which it describes, but it should nevertheless adequately fulfill a predefined modeling purpose. Since a model can only be an approximate description of the original, it is important to be aware of this purpose when constructing models (see Chap. 2 for typical goals in conjunction with DoE). Models used in DoE are polynomials in several variables, in which the y variable is a response, and the x variables are experimental factors. A set of coefficients is used to describe how the y variables depend on the x variables. Often a process or experimental system can only be adequately described by more than one response, in which case there will be one model for each response, and each model will have its own set of coefficients.

Coefficients are estimated from experimental data which are collected in a corresponding experimental design. Estimating model coefficients from experimental data is called model fitting and represents the principle task of statistical analysis to be performed as soon as response values from experiments are available. An equally important task consists of assessing the quality of a fitted model, as an important step towards qualifying it for use in prediction and optimization or whatever other purpose. In fact it is often possible to improve the performance of a model by taking small corrective measures such as those described in Section 5.4.

The important ideas behind empirical modeling are:

- The model must describe the entire behavior of the process or experimental system that is relevant to answering the questions of the experimenter.
- The experimental design is based on the model and determines which information can be extracted from the experimental results. Statistical analysis is only the tool for extracting this information.
- Neglecting experimental design essentially means missing out on finding all the relevant answers. Picking the wrong design also means missing out on relevant answers.
- Enlarging the scope of the questions always means extending the model and adding experiments to the design.

Using models and setting up designs in this fashion requires that one proceed in the systematic way that has been described in Chapter 2. There are several additional aspects that play an important role in modeling:

- Definition of the experimental domain in which the model should be useful for prediction
- Selection of the correct model type
- Choice of the experimental design that corresponds optimally to the experimental domain and to the model type chosen
- Estimation of coefficients by regression analysis
- Qualification and refinement of the model by continued statistical analysis
- Validation of model predictions by confirmatory experiments
- Use of the model for the purpose of finding optimal factor settings

4.2. The Class of Models Used in DoE

The class of models that is normally used in DoE contains only models which are linear with regard to the unknown coefficients. This is why, in order to estimate coefficients, linear regression methods can efficiently be used [32]. Nonlinear or first-principle models, such as mechanical models, reaction-kinetic models, and more general dynamic models are only rarely used directly when designing experiments; they are commonly approximated by simple polynomial models at the cost of restricting the domain of validity of the model. A direct generation of optimal designs for nonlinear models, i.e., models that are nonlinear in the parameters that are to be estimated, is sometimes possible. However, it is particularly important that such models are very accurate, that experimental errors are small and that initial estimates of the parameters are already available. It is this last point that often makes setting up the correct design very difficult [17].

Polynomial models used in DoE are built up as a sum of so-called model terms:

$$y = b_0 + b_A x_A + b_B x_B + b_{AB} x_A x_B + b_{AA} x_A^2 + b_{BB} x_b^2 + \varepsilon$$

This is an example of a quadratic model for two factors A and B, containing a constant term b_0, linear terms $b_A x_A$ and $b_B x_B$, an interaction term $b_{AB} x_A x_B$, quadratic terms $b_{AA} x_A^2$ and $b_{BB} x_B^2$, and an error term ϵ.

The x represent the settings of the factors in the experimental domain. In factorial designs, they are coded as − 1 and + 1. When factors have continuous scales, like temperature or pressure, this coding can be understood as a simple linear transformation of the factor range onto the interval [− 1, + 1]. This transformation is called scaling and centering; the corresponding equation is:

$$x_{\text{centered \& scaled}} = 2\ (x - x_{\text{center}})\ /\ (x_{\max} - x_{\min})\ .$$

Since $x_{\text{center}} = (x_{\max} + x_{\min})/2$, $x_{\max}$ transforms into + 1 and $x_{\min}$ transforms into − 1 (or simply + and − respectively, when using factorial designs). Centering and scaling allows the influences of different factors with different scales to be compared. In the following discussion it is assumed that all factors are either coded or scaled and centered in this fashion.

The b are the coefficients that are estimated by regression after the experiments have been completed. A question that arises is: How do calculated effects in a factorial design as described in Chapter 3 compare to estimated coefficients of linear or interaction terms in a fitted model?

The answer to this question is quite interesting. Estimating coefficients by multiple linear regression and calculating effects for factorial designs as described in Chapter 3 are mathematically equivalent. In fact, coefficients are simply half of the corresponding effects: Calculating a main effect of a factor ME(A) means estimating the difference Δy in the response that has been provoked by changing the factor A from its lower to its upper level. In contrast, the corresponding coefficient, b_A, is the geometrical slope of the curve describing the dependency of y upon x, i.e., $\Delta y/\Delta x$. Since DoE is based on scaled and centered variables, Δx is exactly 2. So b_A can be estimated by ME(A)/2. This is also true for interaction effects: b_{AB} can be estimated by IE(AB)/2. Estimators are often denoted by ˆ, so $\hat{b}_A = \text{IE (AB)}\ /2$. The constant b_0 can be estimated by calculating the mean of all response values.

For the example that was discussed in Section 3.2, the model equation is $y = b_0 + b_A x_A + b_B x_B + b_{AB} x_A x_B$ and the estimated coefficients are:

$$\hat{b}_0 = \tfrac{1}{5}\ (74+92+70+60+62) = 71.6\%$$
$$\hat{b}_A = \text{ME (A)}\ /2 = 5\%$$
$$\hat{b}_b = \text{ME (B)}\ /2 = 11\%$$
$$\hat{b}_{AB} = \text{IE (AB)}\ /2 = 4\%$$

The benefit of using coefficients lies in the greater generality of the response surface models. These allow:

- Prediction of response variables within the experimental region
- Use of quadratic models for modeling maxima and minima
- Use of mixture models for modeling formulations (see Section 7.1)
- Use of dummy variables for modeling categorical factors (see Section 7.2)
- Correcting factor settings when prescribed settings cannot be exactly met in the experiment
- Nonstandard domains for the model (and the design) that are subject to additional constraints

4.3. Standard DoE Models and Corresponding Designs

For standard DoE models the designs can be chosen off the peg. This means that the design structure is predefined and available in the form of a design table (see, e.g., Fig. 11 for half-fraction factorial designs). For nonstandard models an optimal design has to be generated by using a mathematical algorithm (see Section 7.3).

Standard DoE models are

- Linear models (i.e., linear with regard to the factor variables), containing the constant term and linear terms for all factors involved
- Interaction models, which additionally contain interaction terms of the factors involved
- Quadratic models, which, in addition to all interaction terms, contain quadratic terms

Standard models for two and three factors are shown in Table 2, examples of response surfaces are shown in Figure 12.

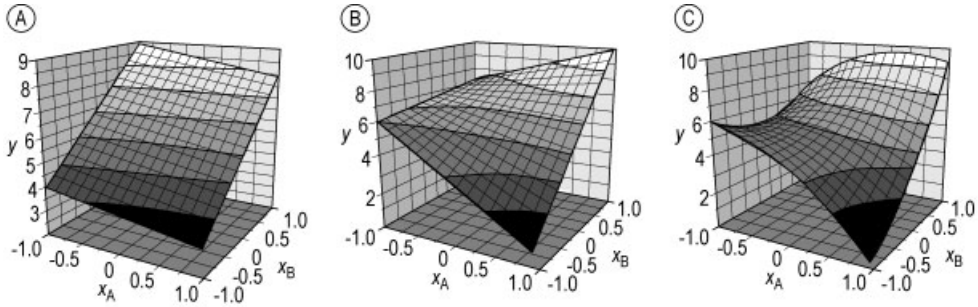

Figure 12. Examples of response surfaces of standard models for two factors: linear (A), interaction (B), and quadratic (C) models

Table 2. Standard models for two and three factors

	Linear model	Interaction model	Quadratic model
Two factors	$y = b_0 + b_A x_A + b_B x_B$	$\ldots + b_{AB} x_A x_B$	$\ldots + b_{AA} x_A^2 + b_{BB} x_B^2$
Three factors	$y = b_0 + b_A x_A + b_B x_B + b_C x_C$	$\ldots + b_{AB} x_A x_B + b_{AC} x_A x_C + b_{BC} x_B x_C$	$\ldots + b_{AA} x_A^2 + b_{BB} x_B^2 + b_{CC} x_C^2$

Adding isolated interaction terms to linear models, taking away interaction terms from interaction models, taking away square terms from quadratic models, or even adding cubic terms like $b_{AAB} x_A^2 x_B$, $b_{ABC} x_A x_B x_C$, or $b_{AAA} x_A^3$, to quadratic models gives nonstandard models which can also be used for DoE. Nonstandard models are also obtained when mixture components are investigated together with normal factors, or when so-called dummy variables (indicator variables) are used to code categorical factors that have three or more settings.

Optimal designs for standard and nonstandard models are:

- Linear models: resolution III factorial designs or so-called Plackett – Burman designs (which are very similar to factorial designs) if no interactions are present
- Interaction models with only some interaction terms: resolution IV factorial designs or so-called D-optimal designs (see Section 7.3); resolution IV designs are also used for linear models when interactions may be present but are assumed to be unimportant
- Interaction models with all interaction terms: resolution V (or higher) factorial designs or D-optimal designs
- Quadratic models: central composite designs (CCD; see below) or so-called Box – Behnken designs
- All nonstandard models: D-optimal designs (see Section 7.3)

Table 3. Design matrix of a CCD for two factors. The run Nos. 1 to 4 form the factorial design, Nos. 5 to 8 the star points, and No. 9 the center point.

No.	A	B	y
1	− 1	− 1	
2	1	− 1	
3	− 1	1	
4	1	1	
5	− 1.41	0	
6	1.41	0	
7	0	− 1.41	
8	0	1.41	
9	0	0	

Factorial designs are discussed in detail in Chapter 3. *Central composite designs* (CCD) are extensions of factorial designs, in which so-called star points and additional replicates at the center points are added to allow estimation of quadratic coefficients (see Fig. 13 and Table 3). The number of star points is simply twice the number of factors and, ideally, the number of replicates at the center is roughly equal to the number of factors. The distance α from the center point to the star points should be greater than one, i.e., the star points should be outside the domain defined by the factorial design. A star distance $\alpha = 1$ is sometimes used; then star points lie in the faces of the factorial design. A good alternative to a CCD is the Box – Behnken design, in which no design points leave the factorial domain [14]. Box – Behnken designs do not contain a classical factorial design as a basis.

CCDs are normally used with full factorial designs, and sometimes with resolution V de-

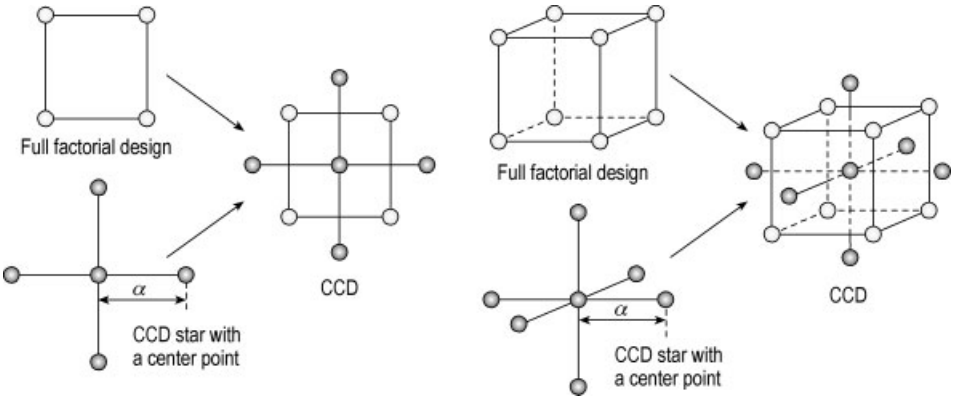

Figure 13. Central composite design for two factors (left) and three factors (right)

signs. Another class of designs, called Hartley designs [33], are similar to CCDs and are based on resolution III factorial designs. When, as is sometimes done in industrial practice, only some quadratic terms are added to an existing interaction model, it is useful to include star points only for those factors for which quadratic terms have been added.

4.4. Using Regression Analysis to Fit Models to Experimental Data

Factorial designs are orthogonal for linear and interaction models which means that coefficients can be estimated independently of each other. This is why calculating effects as described in Chapter 3 is so easy. For more complex models and designs, the analysis will be based on multiple linear regression (MLR) and its variants, such as stepwise regression, variable subset selection (VSS), ridge regression (RR), and partial least squares (PLS) [34]. All of these methods represent ways of fitting the model to the data, in the sense of minimizing the sum of squared distances from measured response values to model values (Fig. 14). They differ in that the minimization procedure is subject to different constraints, and their performance differs only in the case of badly conditioned designs, i.e., when the design is not really adequate for estimating all model coefficients.

If y_i stands for the observed response value at experiment i and $\hat{y}_i$ represents the predicted value at that point, then the least-squares estimates for the coefficients $(b_0, b_A, b_B, b_{AB}, \ldots)$ are those for which $\Sigma(y_i - \hat{y}_i)^2$ is minimized (remember: $\hat{y}_i$ depends on the $\hat{b}$ values).

5. Methods for Assessing, Improving, and Visualizing Models

There are many statistical tools that allow a basic judgement of whether a fitted model is sound. A useful selection of these is:

- The regression measure R^2 to check the quality of fit
- The prediction measure Q^2 to check the potential for prediction and to prevent so-called over-fit, which means that the model is so close to the data that it models experimental errors
- Analysis of variance (ANOVA), to compare the variance explained by the model with variance attributed to experimental errors and to check for significance
- Lack-of-fit test (LoF) to assess the adequacy of the model
- Analysis of the residuals to find structural weaknesses in the model and outliers in the collected data

These methods are used both to qualify models for prediction and optimization purposes and also to find indications of how to improve the models in the sense of increasing their reliability. The different statistical methods are explained and ways to interpret and use them toward model improvement are discussed.

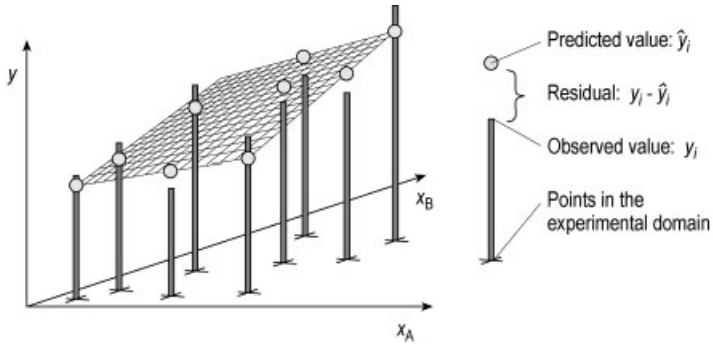

Figure 14. How least-squares regression works: squared distances from the model to the observed values are minimized

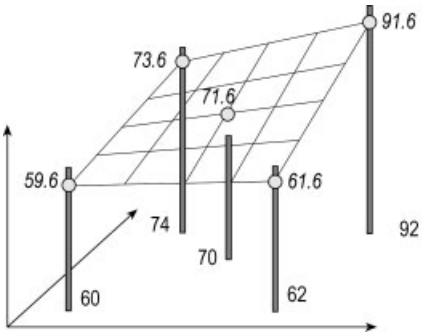

Figure 15. From the example in Section 3.2: observed values y_i and predicted values $\hat{y}_i$ from a model with coefficients $\hat{b}_0 = \bar{y} = 71.6$, $\hat{b}_A = 5$, $\hat{b}_B = 11$, $\hat{b}_{AB} = 4$ (see Section 4.2), so that $\hat{y}_i$, $y_i - \hat{y}_i$ and $y_i - \bar{y}$ can be calculated

5.1. R^2 Regression Measure and Q^2 Prediction Measure

The regression measure R^2 is the quotient of squared deviations due to the model, $SS_{reg} = \Sigma(\hat{y}_i - \bar{y})^2$, and the total sum of squared deviations of the measured data about the mean value $\bar{y}$ $SS_{tot} = \Sigma(y_i - \bar{y})^2$, i.e.,

$$R^2 = SS_{reg}/SS_{tot}.$$

If ordinary least squares is used for fitting the model, $R^2 = 1 - SS_{res}/SS_{tot}$, where $SS_{res} = \Sigma(y_i - \hat{y}_i)^2$, because $SS_{res} = SS_{tot} - SS_{reg}$, the sum of squared residuals. An example for the calculation of R^2 is given in Figure 15 and Table 4. R^2 is always between 0 and 1 and should be as close as possible to 1; how close it should be depends upon the context of the application. When a measuring device is calibrated, R^2 should be above 0.99, whereas when the output of a chemical reaction with several influencing factors is examined, R^2 may well be around 0.7 and still belong to a very useful model.

R^2 is a deceptive measure because it is prone to manipulation: by adding terms to a model, it will always be possible to get a value of R^2 that is almost one, without really improving the quality of the model. On the contrary, many models with a very high R^2 tend to "over-fit" the data, i.e., they model the experimental errors. This is a common and unwanted phenomenon because it decreases the prediction strength of a model. Especially when a design is not orthogonal, one should be wary of over-fit.

To counteract overfit and improve the reliability of model predictions it is useful to consider the prediction measure Q^2. The calculation of Q^2 is similar to that of R^2, except that in the second equation, the prediction error sum of squares (PRESS) is used instead of SS_{res}:

$$Q^2 = 1 - PRESS/SS_{tot}$$

Here, $PRESS = \Sigma\left(y_i - \hat{\hat{y}}_i\right)^2$, where $\hat{\hat{y}}_i$ is the prediction for the ith experiment from a model that has been fitted by using all experiments except this ith one. In a sense $y_i - \hat{\hat{y}}_i$ is a fair measure of prediction errors. PRESS is always greater then SS_{res}, and therefore Q^2 is always smaller then R^2. The relationship between Q^2 and R^2 and an example for the calculation of Q^2 are given in Figure 16 and Table 5.

In good models, Q^2 and R^2 lie close together, and Q^2 should at least be greater than 0.5. This may not be the case if the design is saturated or almost saturated, which means that the number of experiments that have been carried out equals or only slightly exceeds the number of terms in

Table 4. Calculation of R^2 for the example above: SS_{res} = 3.2, SS_{reg} = 648, SS_{tot} = 651.2, hence $R^2 = 1 - SS_{res}/SS_{tot}$ = 0.995. This is a very good value for the regression measure R^2.

	A	B	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$y_i - \bar{y}$
1	–	–	60	59.6	0.4	– 11.6
2	+	–	62	61.6	0.4	– 9.6
3	–	+	74	73.6	0.4	2.4
4	+	+	92	91.6	0.4	20.4
5	0	0	70	71.6	– 1.6	– 1.6
Squared sum					SS_{res} = 3.2	SS_{tot} = 651.2

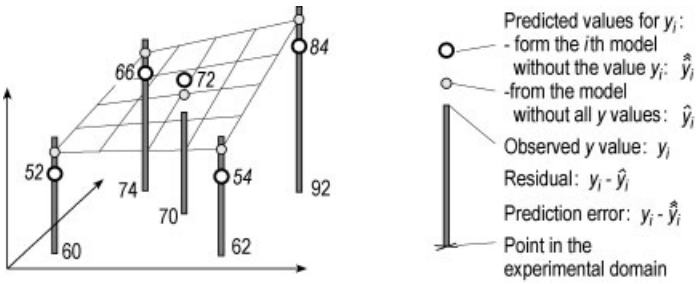

Figure 16. How Q^2 relates to R^2: PRESS $\geq SS_{res}$ and $Q^2 \leq R^2$; $\hat{y}_i$ is usually between y_i and $\hat{\hat{y}}_i$

Table 5. Calculation of Q^2 for the example above: PRESS = 260 and $Q^2 = 1 - 260/651.2 = 0.601$. This is quite a reasonable Q^2 value. Hence, there is no indication of over-fit.

	A	B	y_i	$\hat{\hat{y}}_i$	$y_i - \hat{\hat{y}}_i$	$y_i - \bar{y}$
1	–	–	60	52	8	– 11.6
2	+	–	62	54	8	– 9.6
3	–	+	74	66	8	2.4
4	+	+	92	84	8	20.4
5	0	0	70	72	– 2	– 1.6
Squared sum					260	651.2

the model. In this case, Q^2 may underestimate the quality of the model, because leaving out single measurements may destroy the structure of the design.

It is more dangerous, however, to overestimate the quality of a model, which may happen if many experiments are replicated. In these experiments $\hat{\hat{y}}_i$ will be very close to $\hat{y}_i$ because only one of the measurements is left out in the calculation of $\hat{\hat{y}}_i$. This means that PRESS may be unduly close to SS_{res}, and Q^2 unduly close to R^2. Nevertheless, Q^2 is normally quite a useful measure to prevent overfit.

5.2. ANOVA (Analysis of Variance) and Lack-of-Fit Test

Analysis of variance, usually abbreviated as ANOVA, is a general statistical tool which is used to analyze and compare variability in different data sets. It becomes a powerful tool that can be used for significance testing when assumptions about the error structure underlying the data can be made. In the context of DoE, ANOVA can be used to complement regression analysis and to compare the variability caused by the factors with the variability due to experimental error.

Strictly speaking, “analysis of variance” should be referred to as “analysis of sum of squares” or, even more correctly, “analysis of

the sum of squared deviations", because it is actually this sum of squares that is decomposed (Fig. 17).

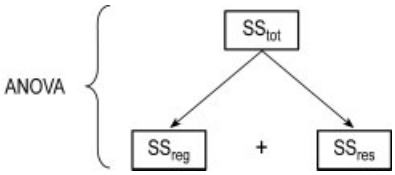

Figure 17. Decomposition of SS_{tot} into SS_{reg} and SS_{res}

However, the aim of ANOVA is to see to what extent the variability in the measured data is explainable by the model and to judge whether the model is statistically significant. To make this comparison, SS_{reg} and SS_{res} must first be made comparable by considering the degrees of freedom.

Already when R^2 is calculated, it is arguable that the number of terms in the model p and the number of runs in the design N should be considered in the calculation, and that $R^2_{adjusted} = 1 - [(N-1)SS_{res}]/[(N-p)SS_{tot}]$ should be used instead of $R^2 = 1 - SS_{res}/SS_{tot}$. But since $R^2_{adjusted}$ normally lies somewhere between R^2 and Q^2, these two are quite sufficient for a first model assessment. In any case, $N - 1$ is the total number of degrees of freedom (of variation with respect to the mean value), $p - 1$ is the number of degrees of freedom of the model (not counting the constant), and $N - p$ is the so-called residual number of degrees of freedom.

ANOVA compares the model to the residuals and tells us whether the model is statistically significant under the assumption that the residuals can be used to estimate the size of random experimental error. $MS_{reg} = SS_{reg}/(p-1)$ is compared to $MS_{res} = SS_{res}/(N-p)$, also called the mean square error (MSE), by subjecting them to a so-called F-test for significance. If the quotient, $F_{emp} = MS_{reg}/MS_{res}$, is greater than 1, then there is reason to suspect that the model is needed in order to explain the variability in the experimental data. If this has to be proven "beyond a reasonable doubt" (i.e., for statistical significance), F_{emp} must be greater than the theoretical value of $F_{crit}(p-1, N-p, \gamma)$, which is always greater than 1, where γ is the desired level of confidence.

Taking the square root of MS_{res} furnishes a reasonable estimate of the standard deviation of a single measurement, provided $N - p$ is not too small and MS_{res} does not change dramatically when terms with small coefficients are added to or removed from the model. This number is called the residual standard deviation (RSD or SD_{res}).

In the example that has been used $SS_{reg} = 648$, $p - 1 = 3$, hence $MS_{reg} = 216$. $SS_{res} = 3.2$, $n - p = 1$, hence $MS_{res} = 3.2$. F_{emp} can be calculated as $F_{emp} = \frac{MS_{reg}}{MS_{res}}$, but a comparison to F_{crit} (3, 1, 95 %) is not meaningful, because there is just one degree of freedom. Also the estimate $RSD = \sqrt{MS_{res}} = 1.789$ should not be taken too seriously in this example.

The lack-of-fit test addresses the question of whether the model may have missed out on some of the systematic variability in the data. This test can only be performed if some of the experimental runs have been replicated: When replicates are present, SS_{res} can be further decomposed into a pure error part $SS_{p.e.}$ and a remaining lack-of-fit part SS_{lof} (Fig. 18).

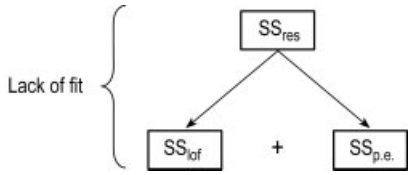

Figure 18. Decomposition of SS_{res} into SS_{lof} and $SS_{p.e.}$

The corresponding significance test compares $MS_{lof} = SS_{lof}/(N-p-r)$ to $MS_{p.e.} = SS_{p.e.}/r$, where r is the total number of replicates of experimental runs in the design. (A replicate count does not include the original run: if there are five runs at the center point, then one of these runs is counted as the original and the four others as replicates; hence, $r = 4$; if a design consisting of eight runs is completely replicated, there are eight runs that have each been replicated once, hence $r = 8$.)

As in ANOVA, a quotient $MS_{lof}/MS_{p.e.}$ greater than 1 means that there may be a lack of fit although the evidence is still weak, whereas a quotient greater than $F_{crit}(N-p-r, r, 1-\alpha) > 1$ means a significant lack of fit at the error probability level α.

The lack-of-fit test method can be regarded as a good complement to the Q^2 prediction measure because the former works well when many replicates are involved, while the latter makes sense when there are only few replicates. Of course, the latter situation is more common in DoE.

5.3. Analysis of Observations and Residuals

Ideally, when models are fitted to data, the model describes all of the deterministic part of the measured values and the part due to experimental error is reflected by the residuals. It is a presumption in linear modeling, i.e., when using least squares fitting as a criterion, that the experimental error is

- Identically (i.e., evenly) distributed over all measurements
- Statistically independent of the measured value, the order of the experiments, the preceding measurement, the settings of the factor variables, etc.

To verify this assumption and to detect outliers, a rudimentary examination of residuals should always be performed as a step toward qualifying the model. The most important tests for residual structure are:

- Test for influence of single observations on the model
- Test for normal distribution and test for outliers
- Test for uniform variance
- Test for independency of measurement errors

These tests can be performed formally as described above for ANOVA and lack-of-fit testing. However, formal testing of hypotheses of the type necessary here usually requires a very high number of residual degrees of freedom $N - p$. Since this number is actively and consciously kept low in DoE, these tests do not often yield useful results. This is why it is common practice to plot observations and residuals in different types of graphs in order to detect structural weaknesses and to find hints on what the problem might be and on how to avoid it.

Common plots are:

- Observed values versus predicted values to see whether there are observations that had undue influence on the model
- Ordered residuals in a normal probability plot to detect outliers and indications of a possibly nonnormal distribution of residuals
- Residuals versus predicted values in order to spot inhomogeneities in variance
- Residuals versus different factor variables to detect weaknesses in the model
- Residuals versus run order to check latent influences of time and autocorrelation

There is a further systematic approach to coping with possible inhomogeneities in variance that was proposed by G. E. P. Box and D. R. Cox in 1964 and that can be summarized in the so-called Box – Cox plot: In addition to fitting a model to the response y, they suggest fitting models to the transformed response $(y^{\lambda} - 1)/\lambda$ for different λ and then to choose $\lambda_{\max}$ such that residuals are closest to normally distributed. The Box – Cox plot, displays performance of a transformation against λ. It is a practical way of finding indications of what type of transformation to the response data may be useful.

5.4. Heuristics for Improving Model Performance

An interesting although somewhat dangerous aspect of DoE and statistical analysis is that of "pruning" models. What is meant by this is the following: When a model displays some weaknesses in the analysis phase explained above, there is often the possibility to improve the performance of a model by simple measures that do not require additional experiments:

- Excluding model terms with insignificant coefficients
- Introducing a transformation
- Excluding observations that seem to unduly dominate the model or that lie far away from the model
- Include one or two terms in the model without overstressing the design, i.e., without having to perform further experiments

Of course there are always situations in which a model remains inadequate and further experiments must be performed to reach the objectives. The following is a set of heuristics, which may help to improve a model in one of several possible situations that may arise during the analysis phase:

– R^2 and Q^2 are both small (i.e., there is just bad fit): check for outliers using a normal probability plot; check that the response values correspond to the factor variables; check pure error if experiments have been repeated; check that no important factors and interactions are missing in the model.
– R^2 is high but Q^2 is very low (below 0.4, i.e., tendency for over-fit): remove very small and insignificant terms from the model (they may reduce the predictive power of the model); check for dominating outliers (by comparing observed and predicted values) and try fitting the model without them (remember, however, that for screening designs with few residual degrees of freedom, Q^2 may be low although the model is good).
– There are clear outliers in the normal probability plot: usually these outliers have not had much influence on the model (otherwise they would be seen when comparing observed and predicted values), however, they often lead to low R^2 values; check what happens when they are removed from the model; check the records, repeat the experiment; mistrust predictions of the model in the vicinity of such outliers; consider that the outlier may contain important information (maybe this is a new and better product).
– There is some structure in a plot showing residuals versus a factor variable (Fig. 19): this is a sign that the model is too weak and should be expanded; this can usually not be done without enlarging the design.

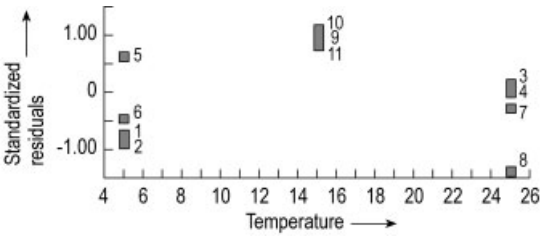

Figure 19. Standardized residuals are residuals divided by RSD = SD_{res} (see Section 5.2). When plotted against a factor variable they may give an indication of how a model can be improved

– In rare cases of analysis it may be observed in a plot displaying residuals versus predicted values that residuals are not homogeneous; this may be the case when a response varies over several orders of magnitude and the size of errors is either proportional to the response values or satisfies some other relation. Indications of this can also be seen in the Box – Cox plot if the optimal exponent $\lambda_{\max}$ deviates significantly from unity. A transformation of the response, e.g., taking logarithms or square roots, may be the correct measure.

Sometimes, the measures suggested above do not lead to a stable model; outliers seem to be present in spite of all attempts to improve the situation, there are bends and curves in the normal probability plot, coefficients are not significant but also not negligible, and so on. This is usually an indication that not enough experiments have been done. It is the authors' recommendation in this case to either

– Reduce all "manipulation" to a minimum (eliminate only obvious outliers from the data and very small coefficients from the model), and use the model knowing that it is weak but may still be useful, or to
– Strip the model down to linear, find the optimal conditions and repeat a larger design here.

In any case the heuristics above cannot repair a model that is simply incorrect, and care must be taken not to succumb to the temptation of systematically perfecting such an incorrect model. This is usually not the purpose of modeling.

5.5. Graphical Visualization of Response Surfaces

When statistical tests such as those described above detect no serious flaws in the model, it is useful to plot the model by reducing it to two or three dimensions and by representing it as a contour diagram (see Fig. 20) or a response surface plot (see Fig. 12).

When the model consists of more than two factors, surplus factors are set constant, usually to their optimum level, i.e., where responses are optimal. Setting a factor constant corresponds to slicing through the experimental domain and reducing its dimension by one. An interesting possibility is setting a third factor constant at three levels, and placing the three contours side by side.

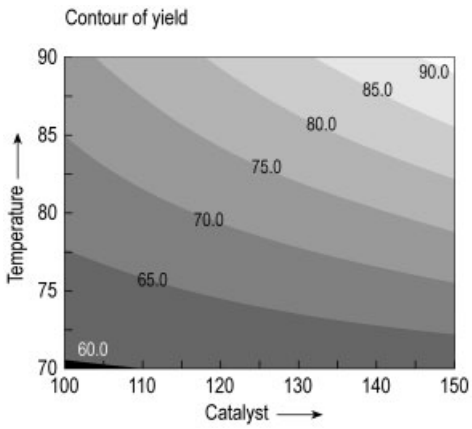

Figure 20. Contour plot of yield, showing the dependency on catalyst and temperature (for the data from Fig. 6 and Fig. 15). It can be seen that there is an interaction between the two factors

Visualizing models is particularly interesting when the modeling involves several responses, because contour diagrams can be used to find specification domains for the factors when specification ranges for the response variables have been imposed (e.g., by the customer of the chemical product to be produced; see Fig. 21).

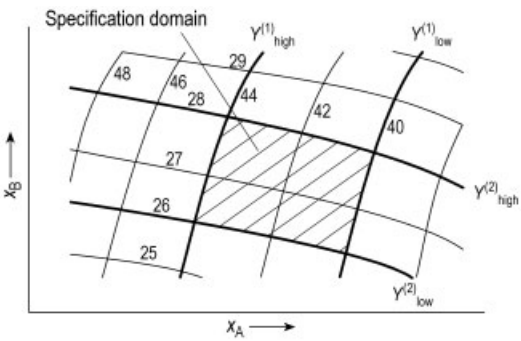

Figure 21. Superimposed contour plot for two responses used for finding specification domains for the two factors x_A and x_B, given the specifications for the responses $Y^{(1)}$ and $Y^{(2)}$

6. Optimization Methods

A major purpose of using response surface modeling techniques is optimization. There are several approaches to optimization depending on the complexity of the situation. Single-response optimizations are easier to handle and are treated in Sections 6.1 and 6.2. Multi-response optimization requires weighting the different responses according to their importance. This can be quite difficult, and the relevant techniques are treated in Section 6.3.

6.1. Basic EVOP Approach Using Factorial Designs

A very basic approach to optimization was proposed in the 1960s by G. E. P. Box and N. R. Draper [15]. It is known as Evolutionary Operation, or EVOP (and has nothing to do with evolution strategy or genetic optimization algorithms). When performing optimization experiments in a running production unit, it is not possible to vary factors over very large domains. Hence, effects will be small and very hard to detect against an underlying natural variability of the response. The idea is to simply repeat a small factorial design for, for instance, two factors, as often as is necessary so that the random noise can be averaged out and the effects become apparent.

The direction of maximal improvement of a response can be deduced from these effects and can subsequently be used to find a new position for a second factorial design. At this new position the second factorial is repeated until significant effects or a satisfactory result is obtained. This procedure can be repeated until a stable optimum is found (Fig. 22).

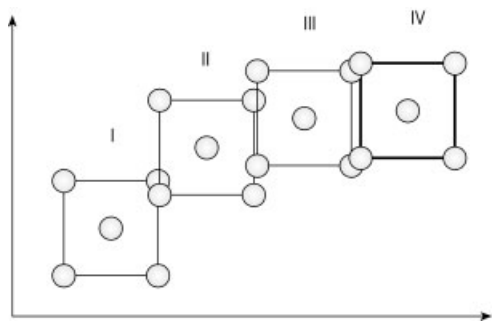

Figure 22. Example of a moving EVOP design (2^2 factorial design with a center point) that has found a stable optimum on the right

6.2. Model-Based Approach

When it is less difficult to establish the significance of a model, as is usually the case on a laboratory scale or in a pilot plant, the model can directly be used for optimization. Gradients of the polynomial models are calculated from the model coefficients because they point to the directions of maximum change in the responses. These directions are then used in an iterative search for a maximum or minimum, as was proposed by Box and Wilson [18].

For quadratic and interaction models, it makes sense to use conjugate gradients, which correct the direction of maximum change by the curvature of the model function (Fig. 23). In this way the search for an optimum can often be accelerated. When a predicted optimum lies within the experimental domain, it is usually quite a good estimate of the real optimum, particularly when a response surface model has been used. When it is outside of the domain, where, by construction, the model is most probably not adequate in describing the process or system under study, it should be validated in the fashion described in Section 6.4.

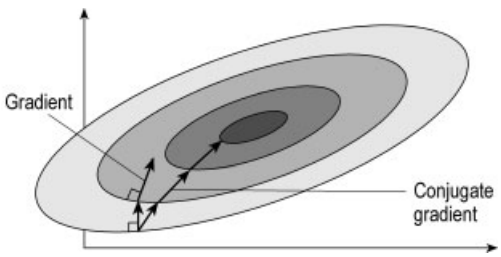

Figure 23. How the conjugate gradient compares to the gradient

6.3. Multi-Response Optimization with Desirability Functions

In an industrial environment, it is usually not sufficient to maximize or minimize one response only. A typical goal is to maximize yields, minimize costs, and reach specification intervals or target values for quality characteristics. But how should one proceed when these goals are contradictory? This question arises quite frequently in practice.

For example, yield, cost, and the quality characteristics are used as responses in an experimental design. The design should be chosen to allow adequate models to be fitted to all responses. For each response, a model will be fitted to its corresponding measured data. Now optimization can be started. Before being able to apply any mathematical optimization algorithm, the goal of the optimization has to be translated from a multidimensional target to a simple maximization or minimization problem. This can effectively be done by using desirability functions. A desirability function $d_j(y^{(j)})$ for one response $y^{(j)}$, measures how desirable each value of this response is. A high value of $d_j(y^{(j)})$ indicates a high desirability of $y^{(j)}$. Many types of desirability functions have been proposed in literature (e.g., [35]). Two-sided desirability functions are used if a target value or a value within specification limits is to be reached (Fig. 24, left), whereas one-sided desirability functions are used if the corresponding response is to be maximized (Fig. 24, right) or minimized.

An example of a two-sided desirability function is:

$$d_j\left(y^{(j)}\right) = 0 \text{ for } y^{(j)} > y^{(j)}_{\max} \; or \; y^{(j)} < y^{(j)}_{\min},$$

$$d_j\left(y^{(j)}\right) = \left(y^{(j)} - y^{(j)}_{\min}\right)^s / \left(y^{(j)}_{\text{target}} - y^{(j)}_{\min}\right)^s$$
$$for \; y^{(j)}_{\text{target}} > y^{(j)} > y^{(j)}_{min},$$

$$d_j\left(y^{(j)}\right) = \left(y^{(j)}_{\max} - y^{(j)}\right)^t / \left(y^{(j)}_{\max} - y^{(j)}_{\text{target}}\right)^t$$
$$for \; y^{(j)}_{\max} > y^{(j)} > y^{(j)}_{target}.$$

This function depends on a target value $y^{(j)}_{\text{target}}$, a minimum acceptable value $y^{(j)}_{\min}$, a maximum acceptable value $y^{(j)}_{\max}$, and two exponents $s,t > 0$ (Fig. 24, left). High values of s and t emphasize the importance of reaching the target; smaller values leave more room to move within $y^{(j)}_{\min}$ and $y^{(j)}_{\max}$.

To build one desirability function for all responses, the geometric mean is usually taken:

$$D(y) = \left[d_1\left(y^{(1)}\right) \cdot d_2\left(y^{(2)}\right) \cdot \ldots \cdot d_k\left(y^{(k)}\right)\right]^{1/k}$$

By substituting the y by the predicted values of the model functions, desirability D becomes a function of the factor variables, $D = D(\boldsymbol{x})$. To find the best possible factor settings, this function D should be maximized by using any efficient mathematical optimization procedure.

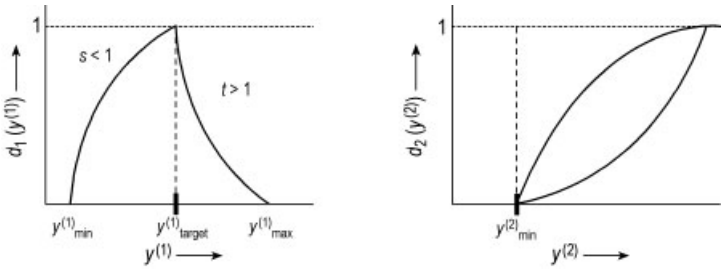

Figure 24. Two responses with corresponding desirability functions: a two-sided desirability function for $y^{(1)}$ (target value $y^{(1)}_{\text{target}}$ desired) and two different one-sided desirability functions for $y^{(2)}$ (maximization of $y^{(2)}$ desired)

In many typical applications, production cost and yield tend to increase simultaneously, and the optimizer, in trying to raise yield and lower cost, reaches an optimum somewhere within the experimental domain. Then it is essential to have good information about the quality of predictions in the interior of the domain. It may be necessary to fit a quadratic model to obtain such good predictions.

In a similar fashion, quality characteristics are often contradictory, e.g., more textile strength means less skin and body comfort; again, the optimizer will find an optimal compromise, which means suboptimal settings for the individual responses within the experimental domain.

Often, the predicted optimum is at the border of the experimental region. It is in fact quite typical that, in cases where linear or interaction models are used and no compromises as described above are necessary, two or more corners of the experimental domain are found to be locally optimal. From a mathematical point of view this seems to present a problem, but in fact it is a very promising situation for further improvement of the product or the process under study. Although the models were made to work inside the experimental region, they often also give very useful information just outside the experimental domain.

To be able to make a compromise between different y, it is sometimes more practical to use desirability functions d_j that are not exactly 0 for $y^{(\mathrm{j})} > y^{(\mathrm{j})}_{\max}$ or $y^{(\mathrm{j})} < y^{(\mathrm{j})}_{\min}$, but only close to 0. This allows the comparison of factor settings for which desirability would otherwise be zero, because some of the y are outside of specification. In any case, constructing desirability functions is quite a delicate task, and it is useful to play around with the desirability function when looking for an optimum.

When extrapolating outside the domain of x, remember that predictions using the model deteriorate when the experimental region is left. This means that predicted optima outside the region must always be validated by further experiments.

6.4. Validation of Predicted Optima

There are several reasons why predicted optima from empirical models should always be validated by further experiments:

- Errors in observed responses propagate into predictions of the model and hence into the position of the predicted optima
- The model may be insufficient to describe the relevant behavior of the experimental system under study and lead to bad predictions of optima
- The desirability function used in multiresponse optimization may not correctly depict all aspects of the real goal
- Predicted optima may lie outside the initial experimental domain, where the quality of the model is doubtful
- Predicted optima may be unsatisfying in that not all responses lie near the desired values or within specified ranges.

It is possible within the scope of linear modeling to calculate confidence intervals for all responses, and hence to get a good idea of the quality of prediction. This calculation of confidence intervals is only correct for qualified models as explained in Chapter 5.

Confidence intervals of this type must be mistrusted if there is a significant lack of fit or if Q^2

is very low. They should also be mistrusted if residuals of observed responses in the vicinity of the predicted optima are very large.

In this case and when predicted optima are outside the experimental domain, further experiments are inevitable. A good strategy for this is to find a sequence of experimental settings for which the predictions gradually improve (according to the model). This sequence will start within the experimental domain and will typically leave it at some point (see Fig. 25). Experiments in this sequence should be carefully performed and checked against predicted results. This procedure will usually lead to very good results. It should be emphasized, however, that these experiments are purely confirmatory in nature, they may lead to better products and better processes but they will not lead to better (fitted) models.

7. Designs for Special Purposes

The design methods that have been discussed until now cover many situations in which the cause and effect relationship between several factors and one or more responses of an experimental system are investigated. However there are cases where further methods are necessary to solve modeling problems. Four characteristic situations are treated in this chapter:

- Designs for modeling and optimizing mixtures are used in product optimization
- Designs for categorical factors may be used for raw product screening and for implementing blocking (as described in Section 2.3)
- So-called optimal or D-optimal designs which are implemented for advanced modeling when nonstandard models or irregular experimental domains are to be investigated
- Robust design techniques have the aim of not only optimizing response values but also response variability.

Further interesting topics, such as

- Nested designs, where the levels of factor B depend on the setting of factor A [16, 24]
- The field of QSAR (quantitative structure activity relationships), which is becoming more and more interesting in conjunction with the screening of active ingredients in medicines
- The whole field of DoE for nonlinear dynamic models involving differential algebraic equations

and others, cannot be treated here. The advanced reader is invited to research on his own, particularly on the latter two themes, which are still in constant motion (both of them incidently make extensive use of D-optimal designs to be shortly discussed in the following).

7.1. Mixture Designs

All designs described above assume that all factor variables can be set and varied independently of each other in an arbitrary manner. This is the typical situation in process optimizations, where factors are technical parameters such as temperatures, pressures, flow rates, and concentrations. However, when the goal is to model the properties of a mixture in order to find optimal ratios for its components, then this factor independence may no longer be presumed. There are many branches of the chemical industry and rubber industries where mixtures are investigated, and there is a whole range of characteristic problems:

- Optimizing the melting temperature of a metal alloy
- Reducing cost in producing paints while keeping quality at least constant
- Optimizing the taste of a fruit punch or the consistency of a yogurt
- Increasing adhesion of an adhesive
- Finding the right consistency of a rubber mixture for car tires

These are cases where experimental design methods must be modified to cope with the fact that all components of the mixture must add up to unity, i.e., for three components, $x_A + x_B + x_C = 1$. Nevertheless there are useful models and designs for mixtures. They are based on the so-called mixture triangle or the mixture simplex when four or more components are involved (Fig. 26).

All simplices have the following properties, which facilitate analysis and the visualization of results:

- Response plots can be generated based on the mixture triangle (as contour plots).

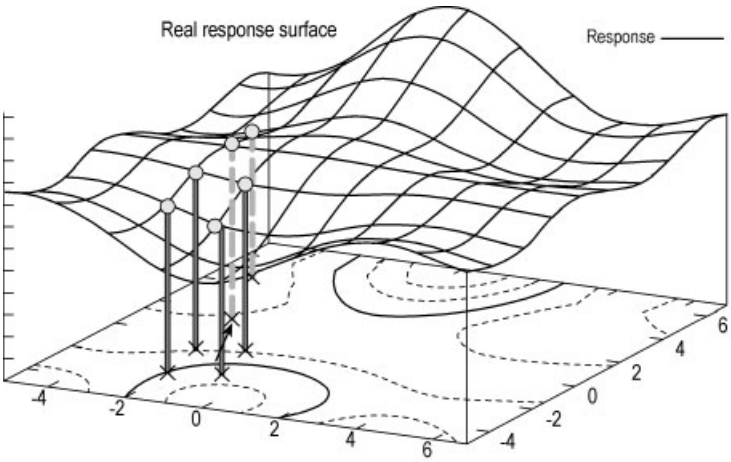

Figure 25. Predicting optima based on the model for increasing domains: A strategy in optimization consists of predicting a series of optima while slowly increasing the domain. Experiments should be performed at these predicted optima for validation purposes. In this way either a satisfactory optimum is reached or a further design is employed

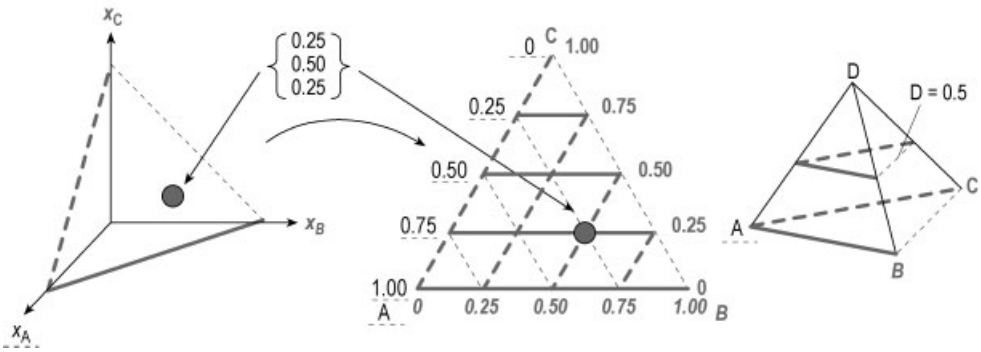

Figure 26. For three-component mixtures the cube is no longer the appropriate geometrical object for modeling the experimental domain. The equation $x_A + x_B + x_C = 1$ defines a triangle. Unfolding this triangle leads to the mixture triangle for three factors (left).
Mixtures with four components can be modeled by a simplex; setting one of the factors constant again leads to a triangle for the others, e.g., $x_D = 0.5$, i.e., $x_A + x_B + x_C = 0.5$ (right)

- A simplex is very similar to the cube in that its boundaries and also its sections parallel to its boundaries are again simplices in a lower dimension. This allows contour plots to be used in the mixture triangle even in situations where more than four components are involved; just set the surplus factors constant and put the three important ones into the mixture triangle.
- Lower and upper levels of the mixture factors can be visualized geometrically as cutting off parts of the simplex. An active upper level means cutting off a corner, an active lower level means cutting off the base (or side) of a simplex. When upper levels are active, the experimental domain is no longer a regular simplex, but only a subset thereof, a so-called irregular mixture region.

The best experimental designs for regular mixture regions are so-called simplex lattice designs. Simplex lattice designs place experiments at the corners and along edges and axes of a simplex. Depending on the complexity of the model that is to be fitted, different strategies in picking out the points are used. The most typical design for a linear model is the axial simplex design which consists of experiments at the corners, at the centroid (equal amounts of all components) and midpoints of the axes from the corners to the centroid (Fig. 27, Table 6). Details can be found in [36].

Table 6. Design matrix for an axial simplex design for three mixture-components

Run no.	Comp. A	Comp. B	Comp. C
1	1	0	0
2	0	1	0
3	0	0	1
4	0.6666	0.1667	0.1667
5	0.1667	0.6666	0.1667
6	0.1667	0.1667	0.6666
7	0.3333	0.3333	0.3334

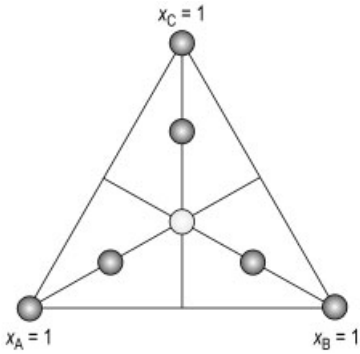

Figure 27. Geometrical view of an axial simplex design for three-component mixtures

When investigating mixtures, it is important to keep in mind that, although it is possible to correlate changes in the responses to changes in the factor settings, it is principally impossible to tell which factors have caused these changes in the responses, because whenever one factor has changed, at least one other factor has also changed. So which one should be made responsible? Example 1: A fruit punch containing pineapple, grapefruit, and orange juice seems to become tastier when pineapple juice is added. Example 2: A long drink containing pineapple and orange juice and vodka seems to become less bitter when pineapple juice is added. Is it possible, in the two examples, to attribute the change in taste to the change in the amount of pineapple?

When mixtures are modeled and optimized, it is not always necessary to use simplex designs and corresponding models. There are principally three ways to proceed, only one of which involves using simplex designs:

- Define component ratios as factors and use the classical factorial design or a CCD (instead of ratios, any other transformation leading to independent pseudocomponents can be used).
- Use a simplex design, as described above, or D-optimal design, as described in Section 7.3.
- Identify a filler component, for example, x_C, which does not have an effect on any of the responses, use a factorial design, a CCD, or a D-optimal design for the remaining components (together with the constraint $x_A + x_B \leq 1$, if necessary), and regression analysis for a model in which the filler variable does not appear.

Analyzing mixture designs requires some care, because due to the dependencies amongst the mixture factors the models must be adapted to the situation. H. Scheffé and D. R. Cox were the pioneers who developed ways to use models correctly for the mixture problem. Details cannot be included here. The reader is referred to the literature [25, 36 – 38].

7.2. Designs for Categorical Factors

Not all factors that influence a product or a process can be quantified. Examples of nonquantifiable or so-called categorical (or discrete) factors are:

- The supplier of a raw material, who may influence some of the quality characteristics of an end product
- The date or the time of the year may influence how well a process will perform
- Different persons doing the same experiments
- The type of catalyst or solvent
- Mutants of a strain of bacteria in a fermentation process

All these are possible influencing factors that cannot be quantified in a satisfactory manner. (In the case of investigating solvents it would be a very good idea to consider polarity as a

quantifiable factor instead of just the type of solvent.) Hence, particularly when more than two instances of a categorical variable must be considered, new design techniques are necessary.

Typical questions that arise in conjunction with categorical or discrete factors are:

- How large is the impact of a qualitative factor? Is it significant?
- If so, which instance yields the best results?
- Does the categorical factor interact with other factors? Are there interactions between different categorical factors?
- Is a possible optimum for the other (continuous) factors robust with respect to varying the categorical factor? If so, where does it lie?

If there are just two instances or categories of a qualitative factor, it is easy to encode them as "−" and "+" and to use factorial designs. The categorical or qualitative factor can then be used in calculations like a continuous quantitative factor. Effects and coefficients can be calculated in the usual way, and they measure how large the influence of changing categories is. The same is true for interactions with other factors; they measure how the influence of these factors on the responses changes when changing categories.

Treating categorical factors with three or more instances is much more delicate. Within generalized linear modeling, it is possible to treat factors with three or more instances by increasing the number of dimensions of the problem. Dummy variables are used to differentiate between instances: Let p_1, p_2 be two dummy variables and encode instances i_1, i_2, i_3 in the fashion shown in Table 7.

Table 7. How to encode three instances of a categorical factor by using two dummy variables

Coding for	p_1	p_2
Instance i_1	1	0
Instance i_2	0	1
Instance i_3	− 1	− 1

Geometrically the three instances become three points in the two-dimensional coordinate system of the dummy variables. The coded design with example values for a response y as well as the corresponding geometrical visualization is shown in Figure 28.

p_1	p_2	y
1	0	6
0	1	8
-1	-1	13

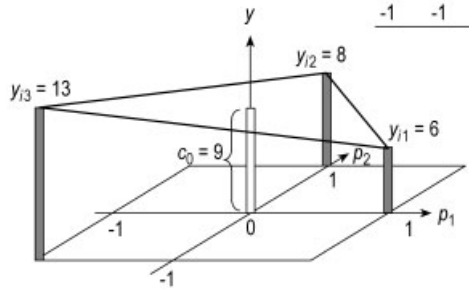

Figure 28. Geometrical and tabular representation of the design and the response values

Now the advantage of using dummy variables becomes apparent: Analysis of data can be done using the same regression methods as described in Section 4.4, and the interpretation of results, although being somewhat different, is still relatively straightforward:

The model has the form

$$y = c_0 + c_1 p_1 + c_2 p_2$$

where c_1 and c_2 are coefficients pertaining to the dummy variables p_1 and p_2.

Model predictions for the instances i_1, i_2, i_3 become:

$$\hat{y}(i_1) = c_0 + c_1$$
$$\hat{y}(i_2) = c_0 + c_2$$
$$\hat{y}(i_3) = c_0 - c_1 - c_2$$

So the coefficients c_1, c_2 are actually a measure of the extent to which instances i_1 and i_2 differ from the mean c_0.

By additionally setting $c_3 = -c_1 - c_2$ the same measure has been found for instance i_3.

For the example shown in Figure 28 the coefficients are $c_0 = 9$, $c_1 = -3$, $c_2 = -1$, $c_3 = 4$.

Designs for categorical variables will normally be adapted to the problem at hand and generated by a computer algorithm. In fact optimal designs, as described in Section 7.3, should be used.

It is interesting to note that dummy variables can be used to model interactions: Just use these variables in interaction terms of a model. The coefficients then measure by how much the effect of another factor varies from a given instance to its mean effect taken over all instances (i.e., its

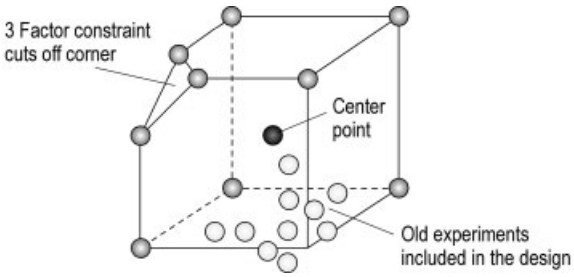

Figure 29. An optimal design with constraints and inclusion of some old experiments

main effect). It would go beyond the scope of this article to go into details about this, in particular about calculation of degrees of freedom. A short digression on dummy variables and their use for modeling blocking can be found in [32].

7.3. Optimal Designs

The usual procedure in planning an experimental design is to identify factors and responses, to determine ranges for the factors, and then to choose a standard design amongst those that have been discussed. This design may be factorial, Plackett – Burman, CCD, Box – Behnken, Simplex, or the like. However, in some situations none of the standard designs are suitable. Such situations typically are:

- Use of nonstandard models
- Use of nonstandard experimental domains (i.e., constraints involving more than one factor)
- Special restrictions on the number of runs or tests that can be performed
- Use of mixture designs when either upper factor levels are active (this means that the regular simplex structure is corrupted) or when normal continuous factors are to be investigated in the same design
- One or more categorical factors with three or more discrete settings are present.

In these cases it is common practice to generate a design by using a mathematical algorithm that maximizes the information that the design will contain. Such designs depend on the model to be used and on the experimental domain to be covered.

Criteria involved in optimizing designs are based on the extended design matrix, usually denoted by $\boldsymbol{X}$, which is built up of the design matrix and extended by a column for each term in the model. The basic idea behind optimal designs is to ensure that there are no correlations between the columns of this $\boldsymbol{X}$ matrix. For if there are correlations here, then the influence of the terms involved cannot be resolved (this situation is similar to that described in Section 7.1 in conjunction with mixture components, which are always correlated).

Working with optimal designs (Fig. 29) involves:

- Defining the experimental domain (including possible constraints)
- Choosing the appropriate model
- Specifying experiments that the experimenter explicitly wants to perform or has already performed
- Selecting the optimization criteria for the design
- Specifying approximately how many experimental runs can be performed.

Typically not just one design is generated, but a whole number of designs of differing size (test or run number). This makes it possible to evaluate designs by comparing the optimality criteria for different designs. Typical criteria are D-optimality, G-optimality, A-optimality, and E-optimality. They are essentially variance-minimizing design criteria in the sense that the variances of model predictions or model coefficients are minimized. The most commonly used criterion is D-optimality, which leads to a maximized determinant of the squared $\boldsymbol{X}$ matrix. For more details, in particular, for further optimality criteria, the reader is referred to special literature on optimal designs [39, 40].

Optimal designs will have a high quality if a sufficient number of experimental runs or tests has been allowed for. They are analyzed by regression analysis like other designs, and the cor-

responding models can be used for prediction and optimization purposes.

7.4. Robust Design as a Tool for Quality Engineering

Sometimes the goal of investigation goes beyond estimating the effects of influencing factors and predicting process behavior or quantifiable product characteristics. A typical application of DoE techniques is toward finding conditions, i.e., factor settings, for which not only quantifiable responses are optimal (usually in a multivariate way as described in Section 6.3) but also the variability of response values is minimal with respect to disturbing or environmental factors that act during production or field use of the product. Typically, these influences can not all be controlled or are too expensive to be controlled even during the experimentation phase.

Doing DoE with the goal of reducing variability is known as robust design. In a robust design, control factors or design factors are varied in a design as described above, typically in a screening design, and experiments are repeated at each trial run so as to estimate the standard deviation (or variance) at each trial. This dispersion measure is employed as a new response value. Statistical analysis and optimization tools such as those described above are then used to quantify and finally to minimize variation, while at the same time improving product characteristics.

Taguchi [1, 19, 20] introduced an additional idea into robust design, namely, the concept of so-called noise factors or environmental factors [1, 11]. These are introduced as perturbing factors in a designed experiment with the idea that they simulate possible external influences which may effect the quality of products after they have left the plant. These factors, which will in fact increase the standard deviation, are varied in a second experimental design which is performed for each of the trial runs in the design for the design factors. To distinguish between the two designs, the design for the environmental factors is sometimes called outer array and that for the design factors is called inner array. The use of outer arrays is recommended, when noise factors can rather cheaply be varied independently of the controlled factors.

Example: During the production of CDs a transparent coating is applied to protect the optical layer. CDs are placed onto a turntable, a droplet of lacquer is applied, and the CD is spun to spread the lacquer. Parameters to be varied are the size of the droplet, the speed and the acceleration of the turntable and the temperature in the room. CDs are to be subjected to different extreme climatic conditions in order to simulate their performance in the field. This is to be done in a climate chamber. Factors to be varied in the corresponding outer array are humidity and temperature in the climate chamber. For the inner array a 2^{4-1} factorial design with 3 realizations at the center point is chosen, for the outer array a simple 2^2 full factorial (Table 8). Results from the four experiments in the outer

Table 8. A design for a quality engineering problem, using a 2^{4-1} factorial with center point as inner array and a 2^2 factorial as an outer array.

Run	Design factors; A, B, C, D				Environmental factors: H, T				Mean	Variance
					H = −	H = −	H = +	H = +		
					T = −	T = +	T = −	T = +		
	A	B	C	D	$y1$	$y2$	$y3$	$y4$	y-mean	y-var
1	−	−	−	−						
2	+	−	−	+						
3	−	+	−	+						
4	+	+	−	−						
5	−	−	+	+						
6	+	−	+	−						
7	−	+	+	−						
8	+	+	+	+						
9	0	0	0	0						
10	0	0	0	0						
11	0	0	0	0						

array are taken together, and the mean and the variance are calculated and subjected to effect calculation or regression analysis.

8. Software

Selected DoE software and suppliers are listed in Table 9.

Table 9. A selection of DoE software tools and providers (as of Nov. 2005).

Tool	Company	WEB link
Design-Expert	Stat-Ease Inc.	http://www.statease.com
D.o.E. FUSION	S-Matrix Corp.	http://www.s-matrix-corp.com
ECHIP	ECHIP Inc.	http://www.echip.com
JMP, SAS/QC	SAS Institute Inc.	http://www.sas.com
MINITAB	Minitab Inc.	http://www.minitab.com
MODDE	Umetrics	http://www.umetrics.com
Starfire, RS/Series	Brooks Automation Inc.	http://www.brooks.com
STATGRAPHICS Plus	Manugistics Inc.	http://www.statgraphics.com
STATISTICA	StatSoft Inc.	http://www.statsoft.com
STAVEX	AICOS Technologies AG	http://www.aicos.com

9. References

1. R. N. Kacker: "Off-Line Quality Control, Parameter Design, and the Taguchi Method (with Discussion)", *J. Quality Technol.* **17** (October 1985) no. 4, 176 – 209.
2. M. S. Phadke: *Robuste Prozesse durch Quality Engineering (Quality Engineering Using Robust Design)*, gfmt, München, 1990 (Prentice Hall, London, 1989).
3. T. Pfeifer: *Qualitätsmanagement: Strategien, Methoden, Techniken*, Hanser, München 1993.
4. J. Wallacher: *Einsatz von Methoden der statistischen Versuchsplanung zur Bestimmung von robusten Faktorkombinationen in der präventiven Qualitätssicherung*, Fortschr.-Ber. VDI Reihe 16 Nr. 70, VDI-Verlag, Düsseldorf 1994.
5. G. E. P. Box, S. Bisgaard: The Scientific Context of Quality Improvement, *Quality Progr.* **June** (1987) 54 – 61.
6. ISO 3534–3: 1999 (E/F): Statistics — Vocabulary and symbols — Part 3: Design of experiments.
7. A. Orth, M. Schottler, O. Wabersky: Statistische Versuchsplanung, Serie: Qualität bei Hoechst, Hoechst AG 1993.
8. S. Soravia: "Quality Engineering mit statistischer Versuchsmethodik", *Chem.-Ing.-Tech.* **68** (1996) no. 1 + 2, 71 – 82.
9. DuPont Quality Management & Technology: *Design of Experiments — A Competitive Advantage*, E. I. du Pont de Nemours and Company 1993.
10. R. A. Fisher: *The Design of Experiments*, 8th ed., Oliver & Boyd, London 1966.
11. S. Bisgaard: Industrial Use of Statistically Designed Experiments: Case Study References and Some Historical Anecdotes, *Quality Eng.* **4** (1992) no. 4, 547 – 562.
12. O. L. Davies (ed.): *The Design and Analysis of Industrial Experiments*, 2nd ed., Oliver & Boyd, London 1956.
13. K. H. Simmrock: "Beispiele für das Auswerten und Planen von Versuchen", *Chem.-Ing.-Tech.* **40** (1968) no. 18, 875 – 883.
14. G. E. P. Box, N. R. Draper: *Empirical Model-Building and Response Surfaces*, John Wiley & Sons, New York 1987.
15. G. E. P. Box, N. R. Draper: *Evolutionary Operation — A Statistical Method for Process Improvement*, John Wiley & Sons, New York 1998.
16. G. E. P. Box, W. G. Hunter, J. S. Hunter: *Statistics for Experimenters: Design, Innovation and Discovery*, 2nd ed; John Wiley & Sons, New York 2005.
17. G. E. P. Box, H. L. Lucas: Design of Experiments in Nonlinear Situations, *Biometrika* **46** (1959) 77 – 90.
18. G. E. P. Box, K. B. Wilson: On the Experimental Attainment of Optimum Conditions, *J. Roy. Statist. Soc.* **B 13** (1951) 1 – 45.
19. G. Taguchi: *System of Experimental Design*, **vols. I and II**, Kraus International Publications, New York 1987.
20. B. Gunter: "A Perspective on the Taguchi Methods", *Quality Progr.* (June 1987) 44 – 52.
21. J. S. Hunter: "Statistical Design Applied to Product Design", *J. Quality Technol.* **17** (October 1985) no. 4, 210 – 221.
22. K. R. Bhote: *Qualität — Der Weg zur Weltspitze (World Class Quality)*, IQM, Großbottwar 1990 (American Management Association, New York 1988).
23. B. Mittmann: "Qualitätsplanung mit den Methoden von Shainin", *Qualität und Zuverlässigkeit (QZ)* **35** (1990) no. 4, 209 – 212.

24. D. C. Montgomery: *Design and Analysis of Experiments*, 6th ed., John Wiley & Sons, New York 2005.
25. E. Scheffler: *Statistische Versuchsplanung und -auswertung — Eine Einführung für Praktiker,* 3., neu bearbeitete und erweiterte Auflage von "Einführung in die Praxis der statistischen Versuchsplanung", Deutscher Verlag für Grundstoffindustrie, Stuttgart 1997.
26. D. E. Coleman, D. C. Montgomery: "A Systematic Approach to Planning for a Designed Industrial Experiment (with Discussion)", *Technometrics* **35** (1993) no. 1, 1 – 27.
27. G. J. Hahn: "Some Things Engineers Should Know About Experimental Design", *J. Quality Technol.* **9** (January 1977) no. 1, 13 – 20.
28. C. D. Hendrix: "What Every Technologist Should Know about Experimental Design", *CHEMTECH* **March** (1979) 167 – 174.
29. R. L. Mason, R. F. Gunst, J. L. Hess: *Statistical Design and Analysis of Experiments with Applications to Engineering and Science*, John Wiley & Sons, New York 2003.
30. E. Spenhoff: *Prozeßsicherheit durch statistische Versuchsplanung in Forschung, Entwicklung und Produktion*, gfmt, München 1991.
31. R. H. Myers, D. C. Montgomery: *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, John Wiley & Sons, New York 2002.
32. N. R. Draper, H. Smith: *Applied Regression Analysis*, 3rd ed., John Wiley & Sons, New York 1998.
33. H. O. Hartley: "Smallest Composite Designs for Quadratic Response Surfaces", *Biometrics* **15** (1959) 611 – 624.
34. I. E. Frank, J. H. Friedman: "A Statistical View of Some Chemometrics Regression Tools", *Technometrics* **35** (May 1993) no. 2, 109 – 148.
35. G. Derringer, R. Suich: "Simultaneous Optimization of Several Response Variables", *J. Quality Technol.* **12** (Oktober 1980) no. 4, 214 – 219.
36. J. A. Cornell: *Experiments with Mixtures — Designs, Models, and the Analysis of Mixture Data*, 3rd ed., John Wiley & Sons, New York 2002.
37. J. Bracht, E. Spenhoff: "Mischungsexperimente in Theorie und Praxis (Teil 1 und 2)" *Qualität und Zuverlässigkeit (QZ)* **39** (1994) no. 12, 1352 – 1360, *Qualität und Zuverlässigkeit (QZ)* **40** (1995) no. 1, 86 – 90.
38. R. D. Snee: "Experimenting with Mixtures", *CHEMTECH* (November 1979) 702 – 710.
39. A. C. Atkinson, A. N. Donev: *Optimum Experimental Designs*, Oxford University Press, Oxford 1992.
40. F. Pukelsheim: *Optimal Design of Experiments*, John Wiley & Sons, New York 1993.

Cyclodextrins

THOMAS WIMMER, Wacker-Chemie GmbH, Burghausen, Germany

1. History 1
2. Structure 1
2.1. Molecular Structure 1
2.2. Secondary Structure 2
3. Properties 2
4. Production 2
4.1. Solvent Process 2
4.2. Nonsolvent Process 4
5. Inclusion Complexes 4
5.1. Preconditions 4
5.2. Preparation 4
6. Derivatives 5
6.1. Ethers 6
6.2. Esters 7
6.3. Other Derivatives 7
7. Uses 8
8. Economic Aspects 9
9. Toxicology 9
10. References 9

1. History

Cyclodextrins were first mentioned under the name "cellulosines" by VILLIERS in 1891. The first detailed description of their preparation and isolation from bacterial cultures was made by SCHARDINGER in 1903, therefore cyclodextrins are sometimes referred to as "Schardinger dextrins". In 1935 PRINGSTON discovered the complexing properties of cyclodextrins. Finally, in the 1950s the cyclic structure of these compounds was elucidated by FREUDENBERG and CRAMER. The first industrial production of cyclodextrins started in the 1970s in Hungary and Japan [1].

2. Structure

2.1. Molecular Structure

Cyclodextrins are cyclic oligosaccharides with a defined number of D-glucose monomers in the ring. All glucopyranose units are linked by glycosidic $\alpha(1 \rightarrow 4)$ bonds.

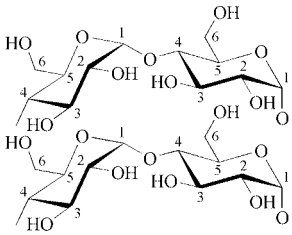

The three most important cyclodextrins (CDs) consist of 6, 7, or 8 glucose units and are named α-cyclodextrin, β-cyclodextrin, and γ-cyclodextrin, respectively. Higher homologues have been described but have gained no industrial significance up to now.

2.2. Secondary Structure

The three-dimensional structure of cyclodextrins resembles a truncated cone with a hydrophobic cavity. All secondary hydroxyl groups at C2 and C3 are directed towards the wider opening of the cavity, whereas the primary hydroxyl groups at C6 are located on the narrow opening. The C-H groups and a ring of glycosidic oxygen bonds are directed inside the cavity

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.e08_e02

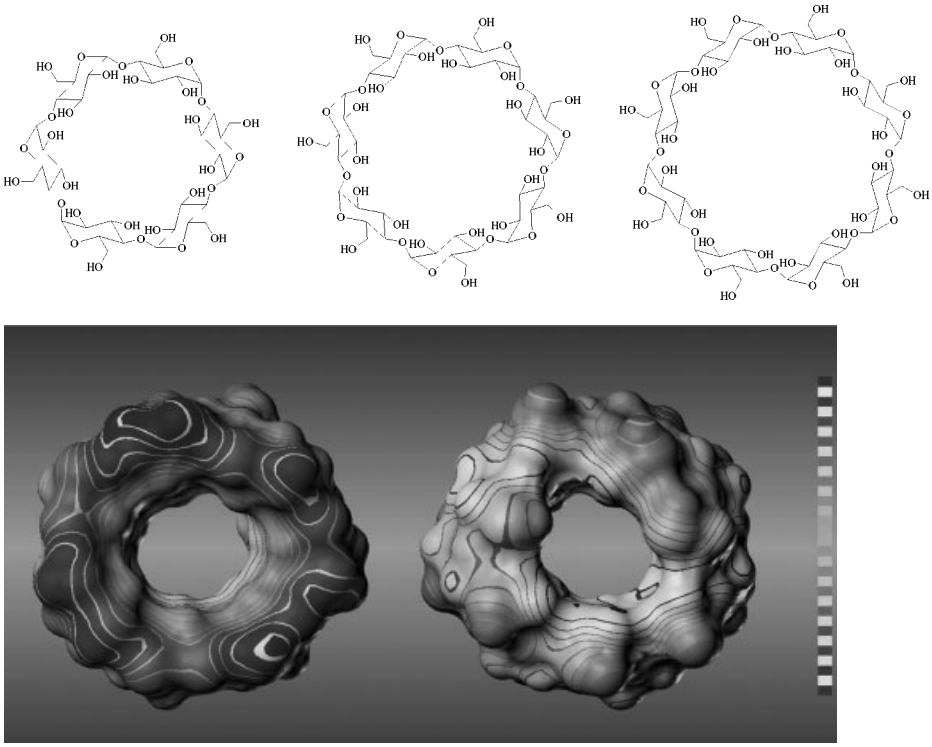

Figure 1. Molecular lipophilicity pattern of β-CD [2]

causing its hydrophobic character. The molecular lipophilicity pattern of β-cyclodextrin is shown in Figure 1.

3. Properties

Physical Properties. Selected physical properties of α-, β- and γ-CD are listed in Table 1.

Cyclodextrins are insoluble in alcohols, ketones, ethers, chlorinated hydrocarbons, and aliphatic and aromatic hydrocarbons.

Chemical Properties [3]. Cyclodextrins are chiral, nonreducing oligosaccharides. Upon oxidation with periodate the glucose rings are cleaved; neither formic acid nor formaldehyde is produced. The only degradation product of all cyclodextrins in acidic solution is glucose. The hydrolysis rate follows the order $\gamma > \beta > \alpha$. Under acidic conditions cyclodextrins are hydrolyzed more slowly than maltooligosaccharides. The glycosidic bond in cyclodextrins is hydrolyzed by α-amylase but not by β-amylase. The rate of enzymatic hydrolysis is fastest with γ-CD, followed by β-CD and α-CD. All cyclodextrins are very stable and highly soluble in alkaline solution (pH > 14). In fact the solubility in water can be highly increased in basic solutions. Under nitrogen atmosphere cyclodextrins are stable up to 250 °C [4].

Substitution of hydrogen of the primary and secondary hydroxyl groups leads to cyclodextrin derivatives. Most reactions are carried out in aqueous solutions (all mentioned in Fig. 4 except acetylations). Other suitable solvents are dimethyl sulfoxide, dimethyl formamide, and pyridine.

4. Production [5]

Cyclodextrins are prepared by enzymatic degradation of starch with cyclodextrin glycosyltransferase (CGTase) at 30 – 90 °C. For industrial production corn and potato starch are most important, but wheat and tapioca starch can also

Table 1. Physical properties of the most important cyclodextrins

	α-Cyclodextrin [*10016-20-3*]	β-Cyclodextrin [*7585-39-9*]	γ-Cyclodextrin [*17465-86-0*]
Formula	$C_{36}H_{60}O_{30}$	$C_{42}H_{70}O_{35}$	$C_{48}H_{80}O_{40}$
Molecular mass	972.85	1135.00	1297.14
Solubility in water (25 °C), g/100 mL	14.5	1.85	23.2
Crystal water, wt %	10.2	13.2 – 14.5	8.13 – 17.7
α_D^{20}, °	+ 148	+ 162	+ 177
mp, °C	> 200 °C	> 200 °C	> 200 °C
pK_a value (25 °C)	12.331	12.202	12.081

be used. Several microorganism are able to produce CGTases, for example *Bacillus macerans*, *Bacillus megaterium*, *Bacillus subtilis*, *Bacillus firmus*, *Bacillus circulans*, *Klebsiella pneumoniae*, or *Klebsiella oxytoca*.

The main steps for the production of cyclodextrins comprise:

- Cultivation of a microorganism which produces the CGTase enzyme
- Separation and purification of the enzyme
- Prehydrolysis of starch with amylases
- Enzymatic conversion of starch with the enzyme
- Inactivation of the enzyme by heat treatment
- Separation of the cyclodextrins from the conversion mixture
- Purification and crystallization from water

4.1. Solvent Process

The aqueous solution of starch is converted to cyclodextrins at a temperature of 30 – 90 °C.

Generally, CGTase enzymes produce all three major types of cyclodextrins, depending on the microbiological source of the enzyme and the conversion conditions. To improve the yield and to alter the ratio of the different cyclodextrins obtained, complexing agents (see Table 2) can be added. Most of the complexing agents form insoluble solid complexes with a certain cyclodextrin. The complex formation shifts the reaction equilibrium towards the desired cyclodextrin and the complex can be separated easily from the liquid conversion mixture. For isolation of the cyclodextrins the complex is dissociated by steam distillation or extraction of the complexing agent with an organic solvent.

The yield can be improved further by utilization of specific enzymes. For example alpha-CGTase produced by *Klebsiella oxytoca* produces predominantly α-cyclodextrin in the initial phase. For large-scale production of γ-cyclodextrin a specific gamma-CGTase was developed.

4.2. Nonsolvent Process

Mainly in Japan a nonsolvent process for the production of cyclodextrins was developed. β-Cyclodextrin produced by this process is approved in Japan as food additive. In this process the reaction mixture consists only of water, starch, and enzyme. The process contains an additional step for the saccharification of the non-converted starch, which is hydrolyzed to maltose and glucose by the action of glucoamylase. The final crystallization of cyclodextrins is difficult in this process.

Only β-cyclodextrin can be produced economically with this technology. Production costs are generally higher than in the solvent process.

5. Inclusion Complexes

The hydrophobic cavity in cyclodextrins is capable to incorporate guest molecules. With this process of molecular encapsulation a cyclodextrin inclusion complex (adduct, clathrate) is formed (see Fig. 3). During the association no covalent or ionic bonds are formed. In solution complex formation and dissociation are in a dynamic equilibrium.

For a complex of 1 : 1 stoichiometry the system can be described as:

$$\mathrm{CD} + \mathrm{G} \underset{\textit{Complex dissociation}}{\overset{\textit{Complex association}}{\rightleftharpoons}} \mathrm{CD} \cdot \mathrm{G}$$

The rate constant k_a for complex association is in the range of $10^0 - 10^8$ mol/s. The complex association therefore is usually a fast process, proceeding within milliseconds. The complex formation is an exothermic reaction ($\Delta H < 0$).

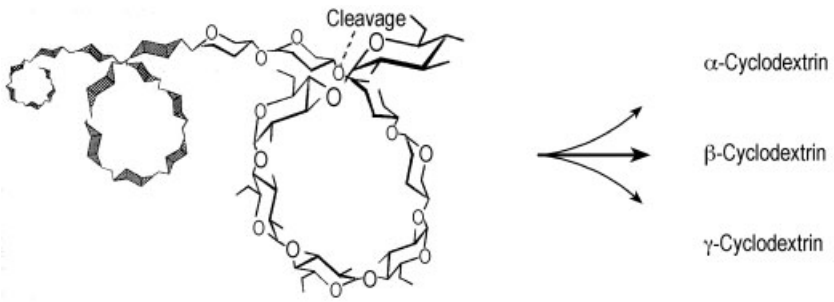

Figure 2. Enzymatic degradation of starch with CGTase

Table 2. Specific complexing agents used for the production of cyclodextrins

α-Cyclodextrin	1-Decanol	Cyclohexane	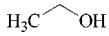 Ethanol
β-Cyclodextrin	Trichloroethylene	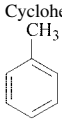 Toluene	
γ-Cyclodextrin	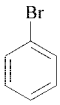 Bromobenzene	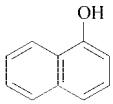 α-Naphthol	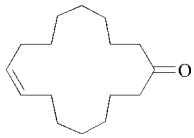 Cyclohexadec-8-en-1-one

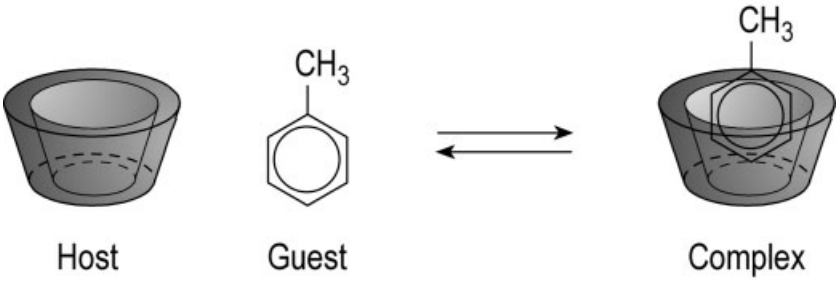

Figure 3. Principles of cyclodextrin inclusion

The complex stability decreases with increasing temperature.

5.1. Preconditions

An important aspects for the formation of cyclodextrin inclusion complexes is the geometry of the guest molecule. The size of a guest molecule must be compatible with the diameter of the cavity, which is dependent on the number of glucose units in the ring. With larger molecules it has to be taken into consideration that only side groups penetrate into the cyclodextrin cavity. In these cases also complexes with higher stoichiometries towards cyclodextrins can be formed. The dependence of the ability to form inclusion complexes with α-, β-, or γ-cyclodextrin on the size of the guest molecules is depicted in Table 3.

The driving forces for complex formation are van der Waals forces, hydrophobic interactions, change in solvation energy for both components and, to a lesser extent, hydrogen bonds. The stability of the complex therefore corresponds to the hydrophobic character of the guest molecule. Very polar or ionic molecules form only weak complexes.

Table 3. Complex formation in dependence on size of the guest molecule

Substance	α-CD	β-CD	γ-CD
Xenon	+	-	-
Chlorine	+	-	-
Bromine	+	+	-
Iodine	+	+	+
Carbon dioxide	+	-	-
Ethylene	+	-	-
Propionic acid	+	-	-
Butyric acid	+	+	-
Cyclohexane	+	+	+
Naphthalene	-	+	+
Anthracene	-	-	+

5.2. Preparation

Most preparations of cyclodextrin complexes are carried out in aqueous systems. Complex formation in the absence of solvent is too slow.

Complexation in Solution. Aqueous solutions of cyclodextrins are agitated with stoichiometric amounts of guest substance at slightly elevated temperatures (40 – 60 °C) for several hours. After the equilibrium is reached water can be removed by freeze drying or spray drying. With many hydrophobic compounds and unmodified cyclodextrins insoluble complexes are obtained which can be isolated by filtration. Drying under vacuum is applied to obtain dry complexes. In some cases the use of polar cosolvents like methanol or ethanol can be of advantage.

Complexation in Suspension. It is not necessary to work with completely dissolved cyclodextrins. Fast complexation can be achieved in stirred aqueous slurries of cyclodextrins and guest substance.

Complexation by Kneading. In the kneading method the water content is further reduced so that the mixture of cyclodextrin, guest, and water forms a paste. To complete complexation this paste is kneaded for 30 – 200 min. Complexation time is usually lower than in the aforementioned methods.

Complexation with Highly Soluble Cyclodextrin Derivatives. Concentrated aqueous solutions of cyclodextrins are stirred with an excess of guest substance. In many cases the formed complexes are very soluble so that uncomplexed material can be removed by filtration. The complex solutions can be used directly or can be spray dried to obtain a water soluble powder.

6. Derivatives

Cyclodextrin derivatives are chemically or enzymatically modified cyclodextrins. As all modifications take place on the hydroxyl groups and all hydroxyl groups are directed towards the outside of the molecule the diameter and the hydrophobicity of the cavity is usually not influenced very much by the modification.

Cyclodextrin derivatives can be divided into the following groups:

$CD–OH \longrightarrow CD–O–R$
R = alkyl, hydroxyalkyl (ether)
R = acyl (ester)
R = glycosyl (branched)
$R = SO_3^-, PO_3^{2-}$
$CD–OH \longrightarrow CD–X$
X = I, Br
$X = NH_2, NHR, NR_2$
X = SH, SR
$CD–CH_2–OH \longrightarrow CD–CO_2H$
Oxidation at C_6

The type of modification can also be classified (see Table 4).

Cyclodextrin derivatives can be characterized by the degree of substitution ($DS = Z/N$). If the substituent can be further substituted, e.g., hydroxyalkylated cyclodextrins, the extend of the substitution can be expressed by the average molar substitution (MS), which indicates the number of substituents per anhydroglucose unit.

The industrially most important cyclodextrin derivatives are the methylated, hydroxypropylated, and branched cyclodextrins (see Fig. 4).

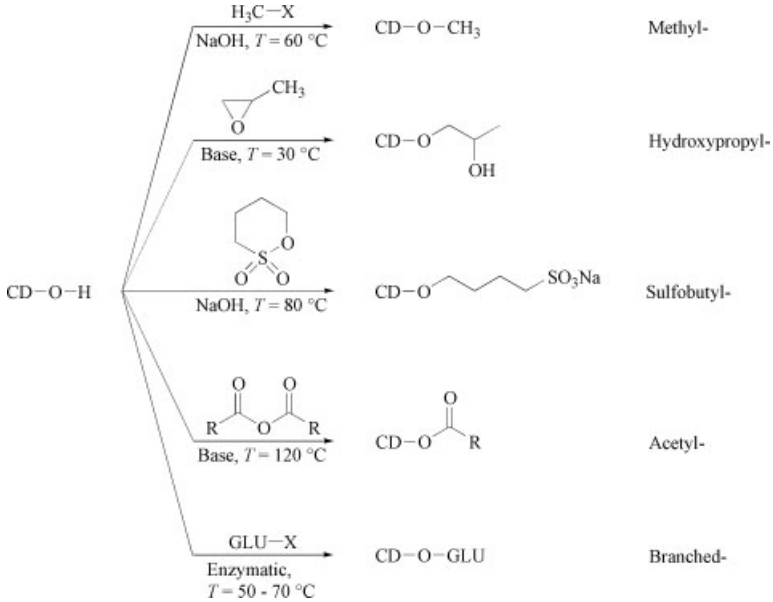

Figure 4. Industrially produced cyclodextrin derivatives

Table 4. Types of modification of cyclodextrins

Type	Z (number of hydroxyl groups/CD which are substituted)	Example
Monosubstituted	$Z = 1$	mono-(6-bromo-6-deoxy)-β-cyclodextrin
Statistically substituted	$1 < Z < 3 \times N*$	methyl-β-cyclodextrin
Persubstituted	$Z = 3 \times N*$	heptakis (2,3,6-tri-O-methyl)-β-cyclodextrin (TRIMEB)
Selectively substituted	$Z = N \times 1{,}2$ or 3	heptakis (2,6-di-O-methyl)-β-cyclodextrin (DIMEB)

$*N$ = number of clucose units (6, 7 or 8)

6.1. Ethers

A comparison of the solubility of different cyclodextrin ethers and natural cyclodextrins in various solvents is given in Table 5.

Methylated cyclodextrins gained interest over the past 20 years. Initial work was focused on specifically substituted heptakis (2,6-di-O-methyl)-β-cyclodextrin and heptakis (2,3,6-tri-O-methyl)-β-cyclodextrin, abbreviated DIMEB and TRIMEB.

Especially DIMEB is very soluble both in cold water (20 – 25 %) and in organic solvents. A major drawback is its poor solubility in hot water and the expensive methods of preparation. The selective substitution in the O2 and O6-positions can only be achieved if the methylation is carried out in the presence of barium salts. Dimethyl sulfate in dimethyl sulfoxide/dimethyl formamide is used as methylating agent.

Highly soluble products (solubility > 50 %) can be obtained economically by statistical substitution with methyl chloride in aqueous sodium hydroxide. A degree of substitution of 1.5 between 2.0 can be obtained in a single-step reaction.

In aqueous solution all methylated cyclodextrins (M-CDs) reduce the surface tension. The viscosity of aqueous M-CD solutions is comparable to that of glucose solutions of the same concentration.

Hydroxyalkylated Cyclodextrins. Among the hydroxyalkylated cyclodextrins only the hydroxypropylated (HP-CD) cyclodextrins are produced commercially [6]. Especially HP-β-CD was developed as a pharmaceutical excipient and for industrial applications. HP-CDs are synthesized by reaction of cyclodextrins with propylene oxide in aqueous solution in the presence of NaOH as catalyst. The desired MS value can be adjusted by the amount of reagent applied and ranges typically between 0.5 and 1.0. The

Table 5. Solubility of cyclodextrin ethers and natural CDs

	α-CD β-CD γ-CD	M-CDs	HP-CDs	CM-CDs
Water	+	++	++	++
Methanol	-	++	++	-
Ethanol	-	++	+	-
Chloroform	-	++	-	-
Acetone	-	++	-	-
Toluene	-	-	-	-

+ soluble
++ highly soluble

lipophilicity of hydroxyalkylated cyclodextrins is lower than that of methylated cyclodextrins.

Ionic cyclodextrin ethers can be obtained by reaction of cyclodextrins with sodium chloroacetate [*3926-62-3*] under alkaline conditions whereby carboxymethyl cyclodextrins (CM-CDs) are formed. Cationic groups are introduced by using 2,3-epoxypropyltrimethylammonium chloride [*3033-77-0*].

Mixed ethers are known but no commercial advantage of their use could be demonstrated up to now.

Table 6. Pharmaceutical formulations with cyclodextrins

Active ingredient	Cyclodextrin	Application/effect
Hydrocortisone	hydroxypropyl-β-CD	increase in solubility
Piroxicam	β-CD	reduction of side effects
Prostaglandin	α-CD	increase in stability
Cefotiam hexetil	α-CD	enhancement of solubility and dissolution rate
Nitroglycerin	β-CD	stabilization
Itraconazol	hydroxypropyl-β-CD	increase in solubility
Chloramphenicol	methyl-β-CD	increase in solubility

6.2. Esters

Partially acetylated cyclodextrins (A-CD) can be produced by reacting a slurry of cyclodextrin in acetic acid with acetic anhydride under sodium acetate catalysis [7]. These esters are very soluble in water (> 30 %) if the DS is in the range of 0.7 – 1.5 and the substitution is random.

With an excess of acetic anhydride fully substituted (DS = 3) products are obtained. These persubstituted cyclodextrins are soluble in polar organic solvents but insoluble in water.

The selectively substituted heptakis (3-acetyl-2,6-di-O-alkyl)-β-cyclodextrins gained some relevance as stationary phase for gas chromatography.

6.3. Other Derivatives

Branched cyclodextrins can be classified in two categories. In *homogeneously branched cyclodextrins* the substituents consists of glucose (G_1) or its oligomers such as maltose (G_2). In *hetero-branched cyclodextrins* the attached groups are other carbohydrates such as galactose.

For the commercial production of G_1-β-cyclodextrin or G_2-β-cyclodextrin highly branched starch is reacted with CGTases to produce a mixture of branched and natural cyclodextrins followed by separation via column chromatography. The $\alpha(1 \rightarrow 6)$ linkage can also be formed by the reaction of excess maltose with β-cyclodextrin in the presence of debranching enzymes like pullulanase.

The solubility of branched cyclodextrins in water is much higher than that of the parent cyclodextrins. Mixtures of branched cyclodextrins are produced and sold in Japan mainly for food applications.

7. Uses

Through the complexation with cyclodextrins the physical properties of the complexed molecules can be changed. The effects of cyclodextrins on guest substances may be summarized as follows:

- Stabilization against light, oxidation, and thermal decomposition
- Solubilization in water
- Controlled release
- Reduction of volatility
- Odor reduction
- Increase in bioavailability

Table 7. Producers and trade names of cyclodextrins

Company	Country	Products	Tradename
Cerestar	USA	β-CD, HP-β-CD , A-CD	Cavitron
Wacker Biochem	USA,	α-CD, β-CD, γ-CD	CAVAMAX
	DE	M-CDs, HP-CDs, A-CDs	CAVASOL
Roquette	FR	β-CD , HP-β-CD	Kleptose
Ensuiko Sugar Refining	JP	β-CD, CD-mixture, branched CD	
Nihon Shokuhin Kako	JP	β-CD, HP-β-CD, CD-mixture	
Janssen Pharmaceuticals	BE	HP-β-CD	Encapsin
Cydex	USA	SBE-β-CD	Captisol
			Advasep

Table 8. Toxicological data of the cyclodextrins

	α-Cyclodextrin	β-Cyclodextrin	γ-Cyclodextrin
Acute toxicology:			
LD_{50} oral, rat	$>$ 10 000 mg/kg	$>$ 12 000 mg/kg	$\gg$ 8000 mg/kg
LD_{50} oral, mouse	$\gg$ 10 000 mg/kg	$>$ 3000 mg/kg	$\gg$ 15 000 mg/kg
LD_{50} i.v., rat	500 – 750 mg/kg	453 – 788 mg/kg	$\approx$ 4000 mg/kg
Subchronic toxicology:			
90 days, oral, rat (NOAEL)[a]	20 % in daily diet	570 mg kg^{-1} d^{-1}	20 % in daily diet
ADI for humans	not allocated (under work)[b]	5 mg kg^{-1} d^{-1}[c]	not specified[d]
Biodegradability (Zahn–Wellens test)	90 % mineralization within 4 days	90 % mineralization within 4 days	90 % mineralization within 4 days

[a] NOAEL = no observable adverse effect level. [b] Submitted to JECFA for evaluation 2001. [c] JECFA decision 1995, published in WHO Food Additive Series 35, 1996. [d] Not specified = JECFA decision 1998: can be used in food up to all quantities technically necessary, published in WHO Food Additive Series 42, 1999.

Pharmaceuticals. Cyclodextrins are useful as excipients in pharmaceutical formulations (see Table 6).

Cosmetics. γ-CD can be applied for odor reduction in self-tanning creams. Retinol, stabilized by γ-CD, is used in anti-wrinkling formulations [8]. An organic peroxide complexed with β-CD is found in tooth pastes [9]. In aqueous skin tonics cyclodextrins are utilized to remove fatty substances from the skin. Perfume oils can be solubilized by HP-β-CD in toiletries [10].

Household. β-CD is used to encapsulate fragrance components in laundry dryer sheets to reduce their volatility [11]. In odor absorbing sprays cyclodextrin derivatives are used to solubilize perfumes and to reduce bad odor [12].

Foods. β-CD is used to remove cholesterol from fatty substances of animal origin like butter or cream [13]. Cyclodextrins can also be utilized in food as carrier for flavors, which can be protected and delivered by molecular encapsulation [14]. It is possible to modify the taste and reduce unpleasant tastes or odors of food with cyclodextrins. Polyunsaturated fatty acids can be stabilized against oxidation by cyclodextrins [15], [16]. In Japan food preservatives like hinokitiol are used as β-CD complex in packaging materials.

Paints and Coatings. Cyclodextrin derivatives are used in the reversible depression of viscosity of associative thickeners [17]. In wood preservation formulations CD derivatives solubilize organic active ingredients like fungicides or insecticides [18].

Chemical Industry. Cyclodextrin derivatives may be applied as phase transfer catalysts for hydrophobic monomers in the synthesis of polymers [19], [20]. In fermentation processes cyclodextrins can be used to "hide" products that are toxic for microorganisms and, thereby enhance the yield.

8. Economic Aspects

Table 7 gives a list of producers and trade names of the different cyclodextrins and their derivatives.

9. Toxicology [21]

Toxicological data of the cyclodextrins are listed in Table 8.

10. References

1. T. Nagai, *Proceedings of the first Int. Symp. on Cyclodextrins* (Eds.: J. Szejtli), Akadémiai Kiadó, Budapest 1982, pp. 15 ff.
2. F. W. Lichtenthaler, S. Immel, *Liebigs Ann. Chem.* (1996) 27 – 37.
3. J. Szejtli (ed.): *Cyclodextrin Technology,* Kluwer Academic Publishers, Dordrecht/Boston/London 1998, 1 – 19.
4. F. Trotta, V. Morone, M. Zanette, G. Camino, *Proceedings of the 9th Int. Symp. on Cyclodextrins*, Santiago de Compostella, Kluwer Academic Publishers, Dordrecht 1999, 27 – 30.
5. G. Schmid in J. Szejtli, T. Osa (eds.): *Comprehensive Supramolecular Chemistry,* **vol. 3,** Elsevier Science, Amsterdam 1996.
6. L. Szente, *New trends in cyclodextrins and derivatives,* Editions de Santé, Chap. 2, D. Duchêne, Paris 1991, 55 – 98.
7. Consortium für elektrochemische Industrie, EP 0 678 525, 1995 (R. Hirsenkorn).
8. Wacker-Chemie, EP 0 867 175, 1998 (J.-P. Moldenhauer, M. Regiert, T. Wimmer).
9. Ausimont, EP 0 895 777, 1999 (U. P. Bianchi, C. Cavallotti, A. Antonini).
10. Siseido, EP 0 366 154, 1990 (H. Matsuda, K. Ito, A. Taki, O. Uejima).
11. Procter & Gamble, US 5 234 610, 1993 (J. M. Gardlik, T. Trinh, T. J. Banks, F. Benvegnu).
12. Procter & Gamble, US 5 668 097, 1997 (T. Trinh et al.).
13. S.A.N. Corman, EP 0 387 708, 1990 (H. Roderbourg, D. Dalemans, R. Bouhon).
14. C. Vaution, M. Hutin, F. Glomot, D. Duchêne, in *Cyclodextrins and their industrial uses,* Editions de Santé, D. Duchêne, Paris 1987, 312 – 320.
15. Staroil, EP 0 470 452, 1992 (T. Bruzzesse, G. Mozzi).
16. Wacker Biochem, EP 1 041 136, 2000 (G. Schmid, M. Harrison, P. Polchinski).
17. Rohm and Haas, US 5 137 571, 1992 (E. K. Eisenhart, R. F. Meritt, E. A. Johnson).
18. F. Imsgard et al., 10th Intern. Symp. Cyclodextrins, Ann Arbor, 2000.
19. Rohm and Haas, US 5 760 129, 1998 (W. Lau).
20. BASF, EP 848 722, 1997 (G. Schornick et al.).
21. G. Antlsperger, G. Schmid, *Proceedings of the 8th Int. Symp. on Cyclodextrins*, Budapest, Kluwer Academic Publishers, Dordrecht 1996, 149 – 155.

Intermetallics

GERHARD SAUTHOFF, Max-Planck-Institut für Eisenforschung, Düsseldorf, Germany

1. Introduction 1
1.1. Definition of Intermetallics 1
1.2. Historical Remarks 2
2. General Considerations 3
2.1. Crystal Structure and Compound Stability 3
2.2. Basic Properties 5
2.3. Criteria for Compound Selection . 6
3. Magnetic Materials 8
3.1. AlNiCo Alloys 8
3.2. FeCo Alloys 8
3.3. Sendust Alloy 9
3.4. Laves Phases 10
3.5. Rare-Earth Compounds 10
4. Superconducting Materials 11
4.1. Laves Phases 11
4.2. A15 Compounds 11
4.2.1. V_3Si 11
4.2.2. V_3Ga 12
4.2.3. Nb_3Sn 12
4.2.4. Nb_3Al 13
4.2.5. Nb_3Si 13
5. Electronic Materials 13
5.1. Silicides 13
5.2. Other Compounds 14
6. Electric Materials 14
7. Thermoelectric and Thermomagnetic Materials 14
7.1. Silicides 14
7.2. Other Thermoelectric Compounds 15
7.3. Thermomagnetic Compounds . . 15
8. Optical Materials 15
9. Hydrogen-Storage Materials . . . 16
9.1. B2 Compounds 16
9.2. Laves Phases 16
9.3. Rare Earth Metal Compounds and Other Phases 17
10. Electrode Materials 17
10.1. Mg_2Ni Alloys 17
10.2. Laves Phase Alloys 17
10.3. RNi_5 Alloys 18
11. Coating Materials 18
11.1. Aluminides 18
11.2. Silicides 19
12. Shape-Memory Materials 19
12.1. Cu-Base Alloys 19
12.1.1. Cu – Zn – Al Shape Memory Alloys 19
12.1.2. Cu – Al – Ni Shape-Memory Alloys 19
12.2. NiAl Alloys 20
12.3. NiTi Alloys 20
13. Medical Biomaterials 21
14. Dental Materials 21
14.1. Amalgams 21
14.2. Cu – Au Alloys 22
15. Wear-Resistant Materials 22
16. Structural High-Temperature Materials 23
16.1. Titanium Aluminide Alloys 23
16.1.1. Ti_3Al Alloys 23
16.1.2. TiAl Alloys 23
16.2. Iron Aluminide Alloys 24
16.2.1. Fe_3Al Alloys 24
16.2.2. FeAl Alloys 25
16.3. Nickel Aluminide Alloys 25
16.3.1. Ni_3Al alloys 25
16.3.2. NiAl Alloys 26
16.4. Silicide Alloys 27
17. References 28

1. Introduction

1.1. Definition of Intermetallics

Intermetallics is the short and summarizing designation for intermetallic compounds and phases, which result from the combination of various metals and which form a tremendously numerous and manifold class of materials. According to the simple definition in [1] intermetallics are compounds of metals whose crystal structures differ from those of the constituent metals, and intermetallic phases and ordered alloys are thus included. During the last 20 years intermetallic compounds (IMCs) have aroused enormous and still-increasing interest in materials science and technology with respect to applications at high temperatures. Various new struc-

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.e14_e01.pub2

tural materials are being developed around the world, in particular in the United States, Japan, and Germany. Besides intermetallics with outstanding high-temperature performance, intermetallics with other outstanding physical properties led to the development of new functional materials much earlier.

The results of the numerous and manifold research and development activities on intermetallics have been published in the respective conference proceedings volumes and review monographs [2 – 8] in addition to scientific journals and research reports.

The present overview, which is an updated summary of a previous report [2], briefly describes the various intermetallics which have already been selected for materials developments or which have been and still are regarded as promising for materials developments.

1.2. Historical Remarks

Intermetallics have been exploited since the beginning of metallurgy, as is exemplified in Table 1. The early intermetallics resulted from alloy systems with low melting temperatures, and the applications relied on the outstanding hardness and wear resistance of intermetallics together with their metallic properties. Because of the metallic behavior of intermetallics, they could be polished, and thus the decorative aspect was also important for many applications. Examples are the bronze coatings in ancient Egypt and the mirrors used by the ancient Chinese, Etruscans, and Romans. Some intermetallics show beautiful colors, e.g. the violet Cu_2Sb ("Regulus of Venus").

Intermetallics became a subject of scientific research with the development of physical metallurgy. Intensive and broad-scale work on intermetallics was initiated by Tammann in Göttingen, Germany, and Kurnakov in St. Petersburg, Russia, around 1900. Work on intermetallics during the first decades of the 1900s already included studies of phase stabilities, phase equilibria, and phase reactions in order to establish phase diagrams, and investigations of chemical, electrochemical properties, and physical properties including magnetism, superconductivity, and mechanical properties. It was recognized that the outstanding hardness of intermetallics is accompanied by an unusual brittleness, and the reasons for this were sought.

In the 1900s intermetallics were used primarily as functional materials. The first industrial applications relied on the special magnetic behavior of certain compounds, and materials developments led, e.g., to Sendust which shows outstanding magnetic properties and wear resistance and is widely used for magnetic heads in tape recorders. In the second half of the twentieth century another important application resulted from the development of new superconducting materials on the basis of the A15 compounds, which are used for superconducting magnets. A third group of functional materials — the shape-memory alloys — make use of a martensitic phase transformation and have found manifold applications since ca. 1970.

An important group of functional materials are the III/V compounds such as InSb, InAs, and GaAs, which have found applications in electronics and thermoelectric power generation. The constituent elements of these compounds represent the transition from metals to metalloids and nonmetals, and the resulting compounds are semiconductors. Hence these intermetallic semiconductors lie outside the field of true intermetallics with metallic behavior.

Intermetallics were not used as structural materials in the past because of their brittleness. The only noteworthy exception is the continuing use of amalgams as dental restoratives. However, various intermetallics have been successfully used as strengthening second phases in conventional alloys for structural applications at high temperatures. Thus it became clear that intermetallics are promising as structural materials for high temperatures because of their high hardness and stability, and numerous investigations began around 1950 with the aim of clarifying the potential of intermetallics for structural applications. Various promising candidates were identified, but the respective materials developments were hampered by unsolved brittleness problems, and these various activities slowly died out in the 1960s.

One offspring of these activities was the development of the electric heating elements based on $MoSi_2$, which rely on the advantageous chemical behavior of this compound, that is, its high oxidation resistance at high temperatures. In the course of superalloy research, sur-

Table 1. Some applications of intermetallics in past and present times [2]

Since ca.	Material or process	Phase	Application
2500 B.C.	cementation	Cu_3As	coating of bronze tools, etc. (Egypt, Anatolia, Britain)
100 B.C.	yellow brass	CuZn	coins, ornamental parts (Rome)
0	high-tin bronze	$Cu_{31}Sn_8$	mirror (China)
600	amalgam	$Ag_2Hg_3 + Sn_6Hg$	dental restorative (China)
1500	amalgam	Cu_4Hg_3	dental restorative (Germany)
1505	amalgam	Sn_8Hg	mirror surface (Venice)
1540	type metal	SbSn	printing
1910	Acutal	$(Cu,Mn)_3Al$	fruit knife (Germany)
1921	Permalloy	Ni_3Fe	high-permeability magnetic alloy
1926	Permendur	FeCo(– 2V)	high permeability magnetic alloy
1931	AlNiCo	NiAl – Fe – Co	permanent magnet material
1935	Sendust	$Fe_3(Si,Al)$	magnetic head material
1938	Cu – Zn – Al, Cu – Al – Ni	CuZn – Al, $(Cu,Ni)_3Al$	shape memory alloys
1950	pack aluminide coating	NiAl, CoAl	surface coating for protection from environment
1956	Kanthal Super, Mosilit	$MoSi_2$	electric heating elements
1961	A15 compound	Nb_3Sn	superconductors
1962	Nitinol	NiTi	shape-memory alloy
1967	Co – Sm magnets	Co_5Sm	permanent magnets

face coatings that protect the coated materials against high-temperature corrosion and which are based primarily on such intermetallics as NiAl and CoAl were developed.

The above-mentioned molybdenum disilicide is a borderline case of an intermetallic, since silicon is not a metal, but a semiconductor. It is known that the combination of silicon with metals gives rise to compounds with metallic properties like $MoSi_2$, as well as compounds with semiconductor properties, that is, silicides represent the transition from intermetallics to compounds of metals and nonmetals. Nevertheless, silicides are traditionally included in the field of intermetallics because of their many similarities with metals.

The first intensive and successful developments of structural materials were based on the titanium aluminides Ti_3Al and TiAl and began in the early 1970s, although the potential of these compounds for high-temperature applications had already been recognized in the 1950s.

Other successful structural materials were based on the nickel aluminide Ni_3Al, which is the strengthening second phase in the superalloys and was also a candidate compound in the 1950s. This development was given impetus by the discovery of the "ductilizing" effect of small additions of boron, which led to major research programs on intermetallics and excited the general and still-increasing interest in intermetallics. These successful developments initiated work on other less common compounds with the aim of lower density and/or higher application temperatures.

2. General Considerations

2.1. Crystal Structure and Compound Stability

Intermetallics form because the strength of bonding between unlike atoms is larger than that between the respective like atoms. Accordingly, intermetallics have particular crystal structures with ordered atom distributions in which atoms are preferentially surrounded by unlike atoms. The known intermetallic compounds are listed with their crystal structures in [9].

The crystal structure of an intermetallic is determined by the strength and character of bonding in the crystal, which depends on the particular electron configuration. The relation between structure type and atomic properties of the constituent atoms, however, is not a simple one, and therefore various criteria have been used for correlating structure type and compound type. The grouping of intermetallics according to simple criteria is not unambiguous, and it cannot be expected that all intermetallics will show metallic bonding like the constituent metals. In the following a brief survey of three important and rather different groups of intermetallics is given to illustrate this.

The *Zintl phases* are formed by metals on the left-hand side and on the right-hand side of

the periodic table of elements and are characterized by completely filled electron orbitals, in particular, by a full octet shell in the normal case. Thus they may be regarded as valence compounds which satisfy the familiar chemical valency rules. The Zintl phases have crystal structures which are characteristic for typical salts, for example, NaTl with cubic B32 structure (*Strukturbericht* designations are used here and in the following for the crystal structures) and Mg_2Si with cubic C1 structure, and therefore ionic bonding is expected. However, all types of bonding — ionic, metallic, and covalent — and mixtures thereof have been found, depending on the particular electron distribution. Hence the Zintl phases may also be regarded as electron compounds whose electron band-structure energy is a large fraction of the total energy and the crystal structure types of which are related to particular valence electron concentrations (average number of valence electrons per atom).

The best-known electron compounds are the *Hume – Rothery phases*, which have the cubic B2 structure (e.g., β-brass at high temperatures and the transition metal aluminides FeAl, NiAl, and CoAl), the complex cubic A13 structure (β-Mn type; e.g., Zn_3Co, Cu_5Si), or the close-packed hexagonal A3 structure (e.g.; Cu_3Ga, Ag_3Al) for the valence electron concentration VEC = 3/2; the complex cubic $D8_2$ structure (γ-brass type) for VEC = 21/13 (e.g.; Cu_5Zn_8, Fe_5Zn_2); and the A3 structure again for VEC = 7/4 (e.g., $CuZn_3$, Ag_5Al_3). The bonding in such intermetallics is not purely metallic, in spite of the metal-like band structure. For example, the bonding in the above-mentioned NiAl is essentially covalent with some metallic character and no ionic component. Recent ab initio calculations showed that, strictly speaking, the B2 aluminides are not Hume – Rothery electron compounds. This means that the Hume – Rothery rules, which relate crystal structure to valence electron concentration, represent a very simplified picture of the bonding situation. Nevertheless, such rules are useful since they describe reality surprisingly well.

Other intermetallics are best characterized by the size ratios and packing schemes of the constituent atoms and are therefore called size-factor compounds, topologically close packed intermetallics, or *Frank – Kaspar phases*. One important group of these intermetallics are the A15 compounds. The best known size-factor compounds are the Laves phases, which form the most numerous group of intermetallics and which crystallize in the closely related hexagonal C14, cubic C15, or hexagonal C36 structures, whereby $MgZn_2$, $MgCu_2$, and $MgNi_2$, respectively, are typical examples. The valence electron concentration, however, is also an important factor here, since it decides between the three crystal structures. These crystal structures are characterized by high symmetry, high atomic coordination, and high density in close analogy to metals, and indeed the *Laves phases* show primarily metallic bonding according to the scarce information available.

Compounds which form by a phase transition with lowering of crystal symmetry from the solid solution of the constituent elements at a transition temperature below the melting temperature are known as *Kurnakov phases*. In the simplest case the phase transition results from an ordering reaction in the solid-solution lattice in which a superlattice structure is established. Examples for compound formation by atomic ordering are the compounds Fe_3Al with $D0_3$ structure, which is formed by two-step ordering from the bcc solid solution, as well as Ni_3Fe with $L1_2$ structure, CuAu with $L1_0$ structure, and Cu_3Au with $L1_2$ structure, the latter three of which form from fcc solid solutions. This atomic ordering with superlattice formation clearly results from stronger bonding between unlike atoms compared with bonding to like atoms. The difference in bonding energy between unlike and like atoms is usually known as interaction energy and corresponds to an ordering energy. Hence, higher crystal stability and higher ordering temperature are expected for a higher interaction energy. The degree of order, which depends on temperature and composition, as well as the order/disorder temperature can be modeled as a function of the interaction energy for a given compound to various degrees of approximation, for example, to a first approximation by the simple Bragg – Williams model or to a better approximation by the cluster variation or the Monte Carlo method. Apart from the degree of approximation, the predictive power of such models with respect to phase-stability and phase-diagram calculations relies on the knowledge of the interaction energy and on whether such interaction energies

are sufficient to characterize the atomic bonding state. Analyses of various transition metal systems have shown that the modeling with the cluster variation or Monte Carlo method gives good agreements with the experimental observations in most cases, whereas in some compounds with bcc structures the simple Bragg – Williams model gave better results than the more sophisticated models.

It follows that single simple atomic parameters like the interaction energy, atomic size, or valence electron concentration have only a limited validity and are generally not sufficient for predicting the stability of a compound as a function of the constituent elements. A better and more general correlation between the atomic properties and the compound crystal structure should be obtained if the particularities of the electron configurations of the constituent atoms are taken into account; these are reflected in the positions in the periodic table of elements. A better correlation has indeed been found, but the resulting structure maps still have their limitations because the changes in electron configuration due to compound formation are not considered.

It must be concluded that the bonding character and crystal structure of an intermetallic compound can be predicted in a significant way only on the basis of quantum-mechanical ab initio calculations on the respective compound. Much progress has been made in this respect, and for various important compounds with not too complicated structures the crystal energy has been calculated as a function of lattice structure in order to determine the compound stability. However, these calculations are very time-consuming even in simple cases and more progress is necessary in the future in order to give guidance to practical materials developments on the basis of multinary compounds with less simple crystal structures.

The intermetallics do not form a homogeneous group of materials. Instead the term intermetallics comprises a huge variety of compounds which differ drastically with respect to bonding, crystal structure, and properties. Therefore, properties of intermetallics cannot be discussed generally for all intermetallics, but can be discussed only by referring to specific groups of intermetallics.

It might be supposed that the particular crystal structure of an intermetallic reflects the character and strength of bonding in a systematic way and that crystal structure would be a good criterion for compound classification. However, this does not mean that intermetallics with the same crystal structure are similar with respect to bonding and properties. For example, the B2 structure is common to the intermetallic NiAl and to the purely ionic salt CsCl, and in the case of the transition-metal disilicides with hexagonal C40 structure, $NbSi_2$ is metallic but $CrSi_2$ is a semiconductor. Recent ab initio calculations showed that atomic coordination and connectivity is more relevant to crystal energy and stability than symmetry. In view of the complexity of the classification of intermetallics, intermetallics are often grouped according to more practical criteria which refer to similarities in behavior.

2.2. Basic Properties

A basic property is the melting temperature since materials parameters which characterize the deformation behavior are well correlated with the melting temperature. Examples are the elastic moduli, which not only control the elastic deformation, but are also important parameters for describing plastic deformation, and the diffusion coefficients, which determine not only the kinetics of phase reactions, but also the kinetics of high-temperature deformation, that is, creep. Furthermore, the melting temperature is intuitively regarded as a measure of compound stability, since it limits the application temperature range.

However, a compound melts when the Gibbs energy of the solid phase, which is a thermodynamic state function and which controls the phase stability, is higher than that of the liquid phase. The Gibbs energy of a phase is given by the enthalpy and the product of entropy and temperature, that is, its temperature dependence is determined primarily by the entropy. Hence, the melting temperature depends on the differences in enthalpy and entropy between the solid and liquid states and is a rather complex function of the bonding strength. Only the enthalpy is directly related to the internal crystal energy, which is determined by the bonding strength.

Nevertheless, the melting temperature correlates surprisingly well with the enthalpy of formation for sufficiently similar compounds.

It may be concluded that the enthalpy of compound formation may be a better parameter for characterizing bonding strength and compound stability and for correlating this with fundamental properties such as elastic moduli. Formation enthalpies have been determined experimentally, and they can be calculated by quantum-mechanical calculations or estimated by more or less empirical approaches.

However, the conventional formation enthalpy values refer to compound formation from the constituent elements at room temperature, that is, from the solid elemental crystals, and thus these formation enthalpies characterize only bonding strength changes during compound formation. For the complete characterization of bonding strength the total compound enthalpy must be considered, which refers to compound formation from the elements in the gaseous state. The total enthalpy of a compound may be estimated by adding the formation enthalpy and the enthalpies of fusion and vaporization of the constituent elements. As an example, data for the Ni – Al system show a good agreement of estimates with experimental data. Since the formation enthalpies are small compared with the total enthalpies, the enthalpies of the constituent elements are the major contributions to the total enthalpies.

The correlation between the total enthalpy of a compound and Young's modulus, which is one of the three elastic moduli, has been checked for various cubic compounds: for the fcc elements Al and Ni (A1), the fcc-ordered Ni_3Al ($L1_2$), the bcc-ordered FeAl, NiAl, and CoAl (B2), and the cubic Laves phases $CaAl_2$, YAl_2, $LaAl_2$, $NbCr_2$, $ZrCo_2$ and $HfCo_2$ (C15). The correlation is quite good for the Laves phases, in that there seems to be a common scatter band, near which the data of some other compounds lie. However, the data points may be subject to appreciable experimental uncertainties, as indicated by data for Ni_3Al. Such uncertainties for Young's modulus may be due to differences in microstructure, that is, texture and artefacts. Apart from this, the data for the B2 compounds show that Young's modulus is not a simple function of the enthalpy of the compound. This must be expected because Young's modulus is a complex function of the potential wells of the atoms, which cannot be characterized by a single parameter. A similar correlation has been found for the activation energy of diffusion.

It must be concluded that there are physically justified correlations between parameters which characterize bonding strength and compound stability and the macroscopic compound behavior. However, such correlations represent complex functional relationships, and hence they are useful only for order-of-magnitude predictions. More quantitative predictions have to consider the character and strength of bonding in a more detailed way, that is, they have to rely on quantum mechanical calculations which are cumbersome and time-consuming.

Recently, fundamental properties, especially elastic moduli, of various simple compounds have been studied by ab initio calculations, and much progress has been made in this field. Thus, a basic understanding of the bonding character of, for example, transition metal aluminides was reached, and it was shown that the degree of directional *d*-electron bonding determines the ideal cleavage strength. However, it has also been shown for B2 transition-metal aluminides that the ductility or brittleness is too complex a function of strength and character of bonding for establishing simple relations between electron distribution parameters and mechanical properties.

2.3. Criteria for Compound Selection

Candidate compounds for materials developments are selected in view of specific applications, that is, they must show certain properties which make them promising for a particular application. For permanent magnets the "energy product" $(BH)_{\max}$ (B = magnetic induction, H = magnetic field strength) is a figure of merit and should reach high values whereas for other functional materials other physical parameters are decisive.

Intermetallic compounds for structural applications must show a sufficient strength at service temperature which also means a sufficient creep resistance. The creep resistance is correlated with the diffusion coefficient and with the shear modulus, and both parameters in turn with the melting temperature. Hence, the candidate com-

pound should have a sufficiently high melting temperature. The limiting temperature for structural applications of conventional metallic alloys is of the order of 75 % of the melting temperature in many cases. The most advanced high-temperature alloys are the Ni-based superalloys, which are used in gas turbines with service temperatures of up to about 1100 °C. Hence compounds with melting temperatures above 1600 °C must be considered if application temperatures above those of superalloys are aimed at. High-melting compounds may form low-melting eutectics that can affect the stability of the compound during processing or service, and therefore such compounds with low-melting eutectics should be avoided.

For many functional and structural applications density is a very important and sometimes decisive parameter. Structural intermetallics for moving components must show a sufficiently high specific strength, which is the ratio of strength to density and has the dimension of length. The specific fracture strength, that is, the rupture length, of advanced superalloys such as the mechanically alloyed oxide dispersion strengthened (ODS) superalloy MA 6000 is of the order of 15 km, which would have to be surpassed by new structural materials. Compounds which contain light elements — e.g. Ti, Al, Si, Mg — may have such a low density that they compare favorably with conventional alloys in spite of lower strength or restricted temperature range.

The main problem of strong intermetallics is their brittleness, which makes processing and application difficult or impossible. However, the brittleness of intermetallics should be less severe than that of ceramics because the atomic bonding of intermetallics is at least partially metallic, whereas it is primarily covalent or ionic in ceramics. This is illustrated by the temperature dependence of yield stress and deformability of the intermetallic NiAl in comparison with that of nickel on the one hand and of silicon and alumina on the other hand. NiAl which is a candidate compound for high-temperature applications softens with rising temperature at about half the melting temperature T_m, which is the range of Si with covalent bonding and which is higher than that of the ceramic Al_2O_3. However, the brittle-to-ductile transition temperature is about 0.4 T_m for NiAl whereas it is above 0.8 T_m for Si and Al_2O_3, that is, the deformation behavior of such an intermetallic may be regarded as intermediate between those of metals and ceramics.

Plastic deformation is more difficult in intermetallics than in metals and conventional metal alloys because of the stronger atomic bonding and the resulting ordered atom distribution, which gives rise to more complex crystal structures. The brittleness of intermetallics increases with decreasing lattice symmetry and increasing unit cell size. Therefore intermetallics with high crystal symmetry — possibly cubic compounds — and small unit cells are preferred for developing new structural materials. Indeed the candidate compounds FeAl, NiAl, Fe_3Al, Ni_3Al, TiAl for development of structural materials have crystal structures which are cubic (B2, $D0_3$ or $L2_1$, $L1_2$, $L'1_2$) or nearly cubic, that is, tetragonally distorted ($L1_0$, $D0_{22}$).

However, even for such selected compounds, brittleness remains a problem, and the causes for this are manifold. A material fractures in a brittle manner if there is no plastic deformation and stress relaxation at the crack tip, that is, the yield stress is higher than the stress for cleavage or fracture. The reason for this may be an insufficient number and/or mobility of dislocations and/or an insufficient number of slip systems. Since these parameters depend not only on crystal symmetry but also on the specific characteristics of the particular compounds, the aim of compound selection and composition variation by alloying must be dislocations with low energy and high mobility and at least five independent deformation modes (i.e. dislocation slip systems and twinning systems), which are necessary for general homogeneous plastic deformation according to the Von Mises criterion. Crystal anisotropy is an important factor in brittleness, and criteria are used which rely on elastic moduli relations. Brittle fracture may be a result of weak grain boundaries and other microstructure heterogeneities leading to localized deformation and stress concentrations. The disproportion of yield stress and fracture stress may also be due to a very low surface energy, which makes cleavage and cracking easy. This may be aggravated by the segregation of impurities, which usually lower surface energies. Impurities, in particular oxygen and hydrogen, may enter the material from the environment, so en-

vironmental embrittlement is an important issue for many intermetallics. A detailed knowledge of the thermodynamics and kinetics of the particular alloy system is necessary for a reduction of the various embrittling effects.

Intermetallics for any application must be corrosion resistant in the operating environment. For high-temperature applications this means oxidation resistance in many cases. Oxidation resistance is provided by the presence of elements, in particular Cr, Al, Si, which can form protective oxide layers. However, chromium oxide is volatile above 1000 °C and silicon oxide may form low-melting silicates. Therefore, aluminides are strongly favored for high-temperature applications. To act as protective layers, the scales must show a sufficient density and adherence, which may be improved by alloying with further elements (macroalloying) and in particular by adding minor amounts of "active" elements such as Ti, Zr, Hf (microalloying). Furthermore, scales must show sufficient long-term stability, which also means mechanical stability, that is, they must tolerate strains, for example, during creep, without being damaged. In the case of insufficient oxidation resistance protective coatings may be applied. However, the thermal stability of such coatings decreases with increasing temperature because of increasing reaction and diffusion rates and therefore intermetallics for applications above 1100 °C should be inherently oxidation-resistant.

Finally, it must already be possible to prepare an intermetallic with sufficient quality in the laboratory, and from a practical point of view this is the basic requirement for any materials development. The processing of high-strength intermetallics is difficult because of their brittleness, and the development of the necessary processing techniques is a very demanding task. It is, however, also a very important task, because a poor quality increases the apparent brittleness and reduces the strength, and this may then preclude any applications of the tested material.

3. Magnetic Materials

3.1. AlNiCo Alloys

The advantageous magnetic properties of Fe – Ni – Al alloys with suitable composition were first discovered by Mishima in 1931, and this initiated the development of the AlNiCo alloys for applications as permanent magnets (→ Magnetic Materials, → Cobalt and Cobalt Compounds, Chap. 6.3). The basic alloy contains about 50 atom % Fe, 25 atom % Ni, and 25 atom % Al, to which various other elements (Co, Cu as macroalloying elements and Nb, Ti as microalloying elements) are added for property optimization. The alloys are single-phase at high temperatures and decompose spinodally on cooling at a controlled rate to form a two-phase microstructure with interlocking phase distribution. The additional alloying elements Co, Mn, Ti, and Cu, which dissolve preferentially in the B2 compound, and Cr, Mo, V, Nb, and Si, which dissolve preferentially in the disordered Fe-rich compound, control the decomposition process of the AlNiCo alloys. The constituent phases of these alloys are the weakly ferromagnetic NiAl base compound with B2 structure and the strongly ferromagnetic, disordered bcc Fe-rich compound. The interlocking phase distribution and the strong shape anisotropy of the Fe-rich compound result in high coercivity. The magnetic behavior is a sensitive function of pretreatment.

AlNiCo alloys, which are produced by casting or sintering, are brittle and hard; they can be machined only by grinding, spark erosion, and electrochemical milling, and they resist atmospheric corrosion well up to 500 °C. The magnetic properties are summarized and compared to those of other alloys in [10 – 12]. AlNiCo alloys are superior to other permanent magnet materials in resisting the effects of temperature on magnetic properties. Presently, NiAl is also considered an attractive candidate compound for other functional applications, for example, contacts in thin-film electronic devices or high-voltage electrodes in electron-optical instruments.

3.2. FeCo Alloys

Permendur, an FeCo alloy containing 2 % V, is a soft magnetic material with high saturation induction (e.g., for electromagnet pole tips [12] (→ Magnetic Materials). These alloys also exhibit a high positive magnetostrictive coefficient, which has made them useful in sonar transducers

and in high-accuracy positioning devices. Similar Fe – Co – V alloys with more V and less Fe are known as Vicalloy alloys.

The B2 compound FeCo is stable only at temperatures below 730 °C and has a broad range of homogeneity. At 730 °C there is an order – disorder transition with 0.2 % change in volume from B2 to A2 (bcc), and a further structure change from A2 to A1 (fcc) occurs at about 985 °C. The ordering reaction proceeds heterogeneously or homogeneously, depending on temperature. The ordering kinetics are affected by alloying additions, and in particular alloying with 2 atom % V reduces the ordering rate significantly and allows the freezing of the high-temperature disordered structure by quenching to low temperatures. The addition of more than 2 % V leads to alloy decomposition, as is indicated by the Fe – Co – V phase diagram. A still larger effect on the ordering kinetics is produced by alloying with Cr. Below the critical temperature of ordering the degree of long-range order at equilibrium increases continuously from 0 at 730 °C to 0.9 at 500 °C and approaches 1 (perfect order) asymptotically at low temperatures.

The mechanical behavior of FeCo has been studied intensively including creep, fatigue, and fracture. At room temperature FeCo is ductile, with 15 % uniform elongation and subsequent necking, in the disordered state, that is, after quenching from above the ordering temperature, whereas in the ordered state, that is, after annealing below the ordering temperature, it shows only little ductility with 5 % uniform elongation and no necking. Hence cold rolling is difficult in the ordered state. Furthermore, FeCo is subject to hydrogen embrittlement at room temperature.

The flow stress increases with decreasing degree of order, that is, with rising temperature, to reach a maximum at about 700 °C just below the ordering temperature when FeCo is only partially ordered. Above the flow stress peak temperature, normal softening is observed due to thermal activation. In other words, FeCo, like other intermetallics, exhibits an anomalous positive temperature dependence of flow stress.

Ordered FeCo – 2V exhibits more homogeneous deformation and a higher strain-hardening rate than disordered FeCo – 2V at all temperatures. The yield stress and fracture stress increase with decreasing grain size according to the usual Hall – Petch relationship. The fatigue resistance is improved by ordering with only a small effect of the environment. The low ductility can be increased by adding Ni as fourth element, which changes the microstructure, and still more by thermomechanical treatment, which results in a Hall – Petch-type dependence of ductility on grain size or subgrain size. Even cold rolling can reduce the brittleness. The brittle-to-ductile transition temperature in the ordered state can be reduced significantly by deviations from stoichiometry, that is, by decreasing the Co content, which decreases the critical temperature of ordering. Positive effects on cold workability have been produced — apart from alloying with V or Cr which both slow down the ordering kinetics — by alloying with V, Nb, Ta, Cr, Mo, W, and C, whereas Be, Ti, Zr, Mn, Cu, Ag, Au, B, Al, and Si are ineffective.

Alloying of Fe and Co with precious metals leads to the compounds FePt, CoPt, FePd with $L1_0$ structure, which is an ordered fcc structure with tetragonal distortion. These compounds have been of importance for applications as permanent-magnetic materials, and in particular CoPt, which is used in the partially ordered state, excels because of isotropy, ductility, easy machinability, and resistance to corrosion and high temperatures (see also [12]). However, CoPt alloy magnets are seldom used now since these alloys have been replaced by rare earth magnets with superior magnetic properties.

3.3. Sendust Alloy

The iron aluminide Fe_3Al with DO_3 structure exhibits an outstanding high magnetic permeability which makes it useful as a magnetic material. The magnetic Fe_3Al-based alloys are known as Alfenol alloys [12]. Fe_3Al is formed on cooling by ordering reactions in the solid state that transform the bcc disordered solid solution, which is stable above about 800 °C, first into the FeAl compound with B2 structure, which is stable between about 800 °C and 550 °C, and then into Fe_3Al with the $D0_3$ structure. The magnetic properties can be improved by partial substitution of Al by Si, which does not change the crystal structure. This gave rise to the development of the Sendust alloy which is based on $Fe_3(Al,Si)$

and which has found widespread application as magnetic head material in recorders.

3.4. Laves Phases

The Laves phases, sometimes designated as Friauf – Laves phases, with AB_2 composition in the binary case form a very large group of intermetallics which crystallize with the hexagonal C14 structure, the cubic C15 structure, or the dihexagonal C36 structure. These structures are topologically close packed (tcp) structures, that is, the Laves phases belong to the family of the Frank – Kaspar phases. The stability of the three crystal structures C14, C15, and C36 is controlled by the atomic-size ratio of the atoms A and B and by the valence electron concentration of the Laves phase.

Various Laves phases have been regarded as promising for both functional and structural applications. Some transition-metal Laves phases, such as TFe_2 (T = Ti, Zr, Hf, Nb, Mo) and $NbCo_2$, as well as the closely related rare-earth Laves phases, show interesting magnetic properties as a result of the particular electronic structures. The rare-earth compounds $TbFe_2$ and $SmFe_2$, both with C15 structure, exhibit huge magnetostrictions at room temperature. In addition there is a magnetoelastic interaction, that is, unusual changes in elastic moduli are observed in a magnetic field. The ternary C15 compound $(Tb_{1-x}Dy_x)Fe_2$, which is known as Terfenol-D, has much potential for applications as a high-power transducer, since it combines a large magnetostriction with a vanishing magnetic anisotropy (for $x \approx \pm\, 0.7$). Like other Laves phases, this material is brittle, with an apparent brittle-to-ductile transition temperature of about 875 °C.

Various rare earth (R) Laves phases, especially RFe_2, can absorb large amounts of hydrogen (see Section 10.2) which may affect their structures, that is, they can undergo structure transformations on hydrogenation, including amorphization, which may be accompanied by decomposition reactions. Such hydrogenation and dehydrogenation reactions have been used for producing fine-grained powders for the powder metallurgical production of so-called giant-magnetostrictive RFe_2 alloys.

3.5. Rare-Earth Compounds

There is a multitude of intermetallic rare earth compounds. The complex crystal structures of these compounds are derived from few basic structures, such as the cubic $D2_1$ structure or the hexagonal $D2_d$ structure, by stacking the structural units in different ways from which the various stoichiometries result. The crystallographic data of most binary, ternary, and quaternary rare earth compounds have been compiled. The rare earth intermetallics are mostly line compounds and have only very narrow composition ranges. The preparation of the compounds is difficult because of the high affinity of the rare earths for oxygen, which leads to reactions with common crucible materials. Various processes have been developed for the preparation of intermetallic rare earth compounds.

The rare earth intermetallics have won a great practical importance because of their outstanding magnetic properties, which depend on composition and crystal structure. $SmCo_5$ and Sm_2Co_{17} led to the development of the rare earth permanent magnets (REPM) which excel by their high energy product $(BH)_{\max}$ and their high coercivity (→ Magnetic Materials). Optimized property spectra have been achieved by substituting other rare earth metals for Sm and Fe and other transition metals and Cu for Co and by appropriate processing. The resulting multinary materials have complex multiphase structures, and the effects of the various structural features on the magnetic properties have not yet been understood completely.

Alloying with the interstitials B, C, and N led to ternary rare earth compounds with very attractive magnetic properties, and outstanding examples are $Sm_2Fe_{17}C_x$, $Sm_2Fe_{17}N_x$, $Nd_2Fe_{14}B$, and $Nd_2Fe_{14}C$. $Nd_2Fe_{14}B$ has become the basis for the development of the Nd – Fe – B permanent magnets, which have an even higher energy product $(BH)_{\max}$ at room temperature than the $SmCo_5$-type and Sm_2Co_{17}-type materials at lower cost; however their behavior at elevated temperatures is less satisfactory. The various REPM materials — $SmCo_5$ type, Sm_2Co_{17} type, $Nd_2Fe_{14}B$ type — are produced primarily by powder metallurgy, though ingot metallurgy and rapid solidification have also been used. The materials are brittle, which makes machining

difficult, and their corrosion resistance is low, which necessitates protective coatings in corrosive environments. $SmCo_5$ suffers an environmental degradation of coercivity due to hydrogen absorption. The hydrogenation of the Nd – Fe – B alloys leads to structure transformation and decomposition reactions and is used for producing fine-grained Nd – Fe – B powders. Because of their most advantageous combination of properties, the rare earth magnets have found a huge market with a broad field of applications. Research and development is rapidly progressing [10, 13].

4. Superconducting Materials

4.1. Laves Phases

A transition to superconductivity at low temperatures has been observed for a large number of Laves phases with C14 or C15 structure. $(Hf,Zr)V_2$ with C15 structure is of particular interest because of its advantageous superconducting properties, since it combines a fairly high critical temperature with a high critical current density and critical magnetic field strength. Processing is difficult because of its poor deformability. Further alloying, in particular with Nb, improves the mechanical behavior to such an extent that multifilament superconductors can be produced which are considered for application in fusion reactor magnets since the superconducting properties are comparatively insensitive to neutron irradiation and mechanical strain. Room-temperature deformation of HfV_2-base alloys occurs primarily by twinning. The Hf – V – Nb phase diagram is available.

4.2. A15 Compounds

The A15 crystal structure is a complex cubic structure derived from the bcc lattice (A2) with eight atoms per unit cell. It is a typical example of the topologically close packed (tcp) structures which result from the stacking of polyhedra of various shapes to accommodate atoms of different sizes. These tcp structures allow a closer packing of atoms of different sizes than the geometrically close packed fcc (A1) and hcp (A3) structures. The A15 structure may be regarded as the basis of the family of compounds with tcp structures which are known as Frank – Kaspar phases.

The compounds with A15 structure form a large and comparatively homogeneous group of intermetallic compounds and are of enormous practical importance since many of them are superconductors (→ Superconductors, Chap. 3.1). Some of them, namely, V_3Ga, V_3Si, Nb_3Al, Nb_3Ga, Nb_3Ge, and Nb_3Sn, exhibit critical temperatures between 15 and 23 K which are only surpassed by those of the ceramic superconductors on the basis of perovskite-type oxides. However, the latter are inferior with respect to critical current density, and the best combination of critical temperature, critical current density, and critical magnetic field strength is found in the A15 superconductors, which have found wide application [12, 14].

The processing of the A15 compounds is difficult because of their brittleness. In the following, various A15 compounds, which have already been applied or have been considered for application, are briefly overviewed.

4.2.1. V_3Si

V_3Si was the first A15 compound discovered to be superconducting, with a critical temperature of 17 K. Stoichiometric V_3Si melts congruently at 1925 °C and is stable at high and low temperatures. A deviation from stoichiometry is possible on the Si-poor side, but reduces the critical temperature due to the resulting constitutional disorder. The critical temperature is also lowered by dissolved hydrogen. Likewise the critical current density and the critical field strength are reduced by elastic tensile or compressive strain. Slightly above the critical temperature, V_3Si transforms martensitically into a weakly tetragonal, still superconducting compound with twin-related thin lamellae. Prior plastic deformation with large strains at high temperatures stabilizes the tetragonal martensitic phase even at temperatures above the martensite formation temperature. The martensitic transformation gives rise to superelasticity and a shape-memory effect.

V_3Si has a high Young's modulus of 213 GPa and an apparent brittle-to-ductile transition temperature (BDTT) of the order of 1200 °C. The

plastic deformation behavior above the BDTT has been studied in detail. The creep behavior is characterized by a comparatively high activation energy in the range of 2 – 11 eV per atom that depends on stress and composition. Creep can be enhanced by superimposed electric currents.

The transition from superconduction to normal conduction in A15 superconductors occurs by the penetration of small fluxoids, that is, quantized magnetic flux vortices, which is characteristic for the type II high-field superconductors. These fluxoids can be pinned by microstructural inhomogeneities, in particular grain boundaries and precipitated particles, and this is beneficial with respect to critical current density and field strength. Consequently, microstructural changes by prior plastic straining with or without recrystallization affect the superconducting properties. The thin lamellae, which are produced by the above-mentioned martensitic transformation, are less effective with respect to flux pinning than grain boundaries, so that their effect on the superconducting properties becomes relevant only in single crystals.

V_3Si is of interest for applications because of its favorable superconducting properties and its high stability. Apart from hot extrusion, V_3Si can be deformed plastically at room temperature in spite of its brittleness by superimposing a high hydrostatic pressure, which makes the production of multifilament superconducting wires feasible for applications in solenoid magnets. Laminar V_3Si can be prepared by solid-state thin-film reactions between V and SiO_2 on Si substrates. However, Nb_3Sn has been found more favorable with respect to superconducting properties, mechanical behavior and processing, and this has precluded the application of V_3Si.

4.2.2. V_3Ga

V_3Ga with A15 structure is stable only up to 1300 °C where it transforms congruently into a V – Ga disordered bcc solid solution. There are indications for a martensitic transformation at very low temperatures, in close analogy to V_3Si, although the conditions for martensite formation are not yet clear for V_3Ga. It has a broad range of homogeneity centered around the stoichiometric composition, that is, it is very stable with respect to deviations from stoichiometry. Because of this stability, which is advantageous for fabrication, and the high critical temperature of about 16 K, V_3Ga is of interest for applications.

The superconducting properties and the deformation behavior of V_3Ga are similar to those of V_3Si. The superconduction transition can be affected by prior plastic deformation. V_3Ga is ductile in compression above 1000 °C and the creep behavior has been studied in detail. As in the case of V_3Si, V_3Ga can be deformed plastically by hot extrusion and at room temperature by superimposing a high hydrostatic pressure which makes the production of multifilament superconducting wires feasible for applications in solenoid magnets. Hybrid high-field magnets with V_3Ga tape coils have been produced successfully. However, Nb_3Sn has been found more favorable with respect to superconducting properties, mechanical behavior, and processing, and this has precluded the commercialization of V_3Ga up to now.

4.2.3. Nb_3Sn

Nb_3Sn with A15 structure forms peritectically at 2130 °C from the liquid solution and solid Nb and is stable only above 775 °C where it decomposes eutectically. The decomposition reaction can be suppressed by not too slow cooling, so that Nb_3Sn with A15 structure is maintained at lower temperatures in a metastable condition. Nb_3Sn transforms into a twinned tetragonal martensite at 43 K, and the transition to superconductivity occurs at about 18 K. This critical temperature is lowered by deviations from stoichiometry, which affect the atomic order, and by elastic strain. Small increases in the critical temperature are possible by addition of ternary alloying components. The transition temperature does not depend sensitively on prior plastic deformation.

Like the other A15 compounds, Nb_3Sn can be deformed readily at high temperatures, e.g., by hot extrusion, whereas deformation at low temperatures is possible only by superimposing a high hydrostatic pressure. The high temperature deformation and creep behavior has been studied in detail, and a beneficial effect of grain size reduction on the brittle-to-ductile transition temperature (BDTT) has been found.

Nb_3Sn superconductors with multifilament geometry, which are needed for electromagnet applications, are easily fabricated by the so-called bronze process which has lead to the widespread commercial application of Nb_3Sn. In the bronze process Nb rods are inserted in a bronze rod or tube which is then processed to obtain a multifilament wire. After wire processing, Nb_3Sn shells are formed around the Nb cores by reaction heat treatments. The final multifilament composite superconductor consists of Nb_3Sn fibers with Nb cores which are embedded in a bronze matrix; the latter is surrounded by a coating of Nb, which serves as a diffusion barrier, and an outer Cu tube, which serves as a stabilizer and carries the current in the case of transition to normal conduction. Various specific processing routes, including powder metallurgy, have been developed for optimized superconductor production, and hybrid superconductor coils are commercially available.

The superconducting properties of the composite superconductor as well as the mechanical properties depend on the particular microstructure which is a function of the various prior processing steps. Much work has been directed at the details of the various processes, e.g. the effects of alloying additions, the diffusion of Sn during Nb_3Sn formation, the grain size dependence of the flux pinning force and the tensile strength and fracture at room temperature, or the effects of the hardnesses of the constituent phases on the composite workability. Recent new developments rely on the beneficial effects of additions of Ti, Hf, Ta, and/or Ge on the superconducting properties.

4.2.4. Nb_3Al

Nb_3Al with A15 structure forms peritectically at 2060 °C and is stable below this temperature in the whole temperature range only with Al-deficient off-stoichiometric compositions. The stoichiometric composition corresponds to the maximum solubility of Al in Nb_3Al and is reached only at 1940 °C in eutectic equilibrium with the liquid and the σ-phase Nb_2Al. The transition to superconduction occurs at 19 K and there is no martensitic transformation. The transition temperature of Nb_3Al is affected by deviations from stoichiometry and perfect order like the other A15 compounds, but in a less sensitive way. The A15 compound with the highest transition temperature is Nb_3Ge, which can be alloyed with Nb_3Al to form the ternary A15 compound $Nb_3(Al_{0.7}Ge_{0.3})$ with the highest critical field strength.

Nb_3Al has long been considered for electromagnet applications because of its advantageous superconducting properties. However, the processing of Nb_3Al is very difficult because of its brittleness of Nb_3Al. Powder metallurgy has been used successfully for preparing Nb_3Al by reaction synthesis and producing, e.g., a Nb_3Al – Ag composite superconductor. Nb_3Al tapes have been prepared by rapid solidification. A novel process with thin Nb – Al sandwiches as starting material has allowed the fabrication of Nb_3Al superconducting wires with good workability. Furthermore various other processes have been used successfully for producing Nb_3Al wires.

4.2.5. Nb_3Si

Nb_3Si is a high-temperature line compound which forms with a tetragonal crystal structure by a peritectic reaction at 1980 °C and is stable only above 1770 °C. Rapid solidification of the liquid produces amorphous Nb_3Si. The A15 structure is unstable and is obtained only by special rapid solidification techniques and by crystallization of amorphous Nb_3Si. Nb_3Si with A15 structure is of interest because a high transition temperature for superconduction is expected, and indeed a transition temperature of 19 K was reported for near-stoichiometric Nb_3Si with A15 structure, which was prepared by crystallization of amorphous Nb_3Si under high pressure.

5. Electronic Materials

5.1. Silicides

Metal(M)-like M_2Si silicides, as well as other transition metal silicides (→ Silicon, Chap. 5), are of major importance for applications as thin-film materials in electronic devices, for example, as low-resistivity interconnects and gates, Ohmic contacts, and Schottky barriers in very large scale integrated (VLSI) circuits [16]. The

most eminent example is Pd_2Si which is used for shallow contacts with very small controlled thickness and extension. The kinetics of thin-film formation and growth are controlled by diffusion and can be optimized by alloying with further elements. The film-formation process affects the distribution of dopants in the Si substrate. Other M_2Si compounds have also been studied with respect to thin-film formation, e.g., Ni_2Si, Co_2Si,and Mg_2Si. Thin oxide layers may form on the silicide films by air exposure at room temperature, which can be a problem for Ohmic contacts or a necessity for a tunneling device.

The transition metal silicides CrSi, MnSi, FeSi, and CoSi crystallize with the cubic B20 structure, whereas NiSi and the noble metal silicides PtSi, IrSi, and PdSi crystallize with the orthorhombic B31 structure. Such MSi compounds, in particular PtSi and also NiSi, are also important thin-film materials for applications in electronic devices.

The disilicides $CrSi_2$, VSi_2, $NbSi_2$, and $TaSi_2$ crystallize with the hexagonal C40 structure. VSi_2 is of interest for applications in electronic devices. $TaSi_2$ has been successfully used in producing microprocessors and other electronic devices. $CrSi_2$ is of interest for applications as thin-film interconnectors. $CoSi_2$ with the comparatively simple cubic C1 structure is applied in large-scale integrated electronic devices. $NiSi_2$ is another compound of primary interest for applications in electronic thin-film devices. $MoSi_2$ is advantageous for applications in electronic devices either as substrate for Si films or as thin-film interconnectors in VLSI circuits. In view of these applications, the conditions for thin-film formation as well as for the formation of thin oxide layers have been studied in detail.

5.2. Other Compounds

Various aluminides, such as Al_3Ti, Al_3Hf, Al_3Ta, and Al_7Cr, are used as shunting layers in integrated circuits to improve the resistance to failure of Al-based interconnects by electromigration stressing [17].

GaAs, InP, InSb, and InAs are compounds with high electron mobilities and high saturation velocities relative to Si and hence are very attractive materials for application in high-speed devices [18]. Such devices are the GaAs metal Schottky field effect transistors (MESFETs), the high electron mobility transistors (HEMTs), the heterojunction bipolar transistors (HBTs), the permeable-base transistors (PBTs), the static inductance transistors (SITs), and the metal insulator semiconductor field effect transistors (MISFETs).

6. Electric Materials

The transition metal disilicides MSi_2 (M = Ti, Zr, Hf, V, Nb, Ta, Cr, Mo, W (→ Silicon, Chap. 5) crystallize with the tetragonal $C11_b$ structure, the hexagonal C40 structure, or the orthorhombic C49 and C54 structures and show extended mutual solid solubilities. These crystal structures are closely related since they only differ in the stacking of the same structural elements.

$MoSi_2$ with $C11_b$ structure has found widespread application as heating elements in high-temperature furnaces at up to 1700 °C because of its advantageous electrical properties and its excellent oxidation resistance. Well-known examples are Kanthal Super and Mosilit heaters. The high oxidation resistance of these materials, which is not shown by the other Mo-rich silicides, is due to the formation of a protective vitreous highly adherent SiO_2 film with prior evaporation of the volatile Mo oxide. However, $MoSi_2$ is subject to oxidative disintegration by pest oxidation at intermediate temperatures between about 300 °C and 600 °C, which occurs at grain boundaries and microstructural defects and thus can be minimized by avoiding these or by alloying with small amounts of Fe and Re.

7. Thermoelectric and Thermomagnetic Materials

7.1. Silicides

The dimagnesium monosilicide Mg_2Si is a semiconductor with a high thermoelectric power and a low thermal conductivity which is advantageous for thermoelectric power generation. The thermal conductivity can be minimized by alloying with Ge to form the ternary $Mg_2(Si,Ge)$ compound. Doping with Ag and Sb produces *p*-

and *n*-type semiconducting $Mg_2(Si,Ge)$, respectively, and the combination of both results in a thermocouple with a high efficiency for thermoelectric energy conversion. The interactions between doping, point defects, and microstructure has been studied in detail.

$FeSi_2$ is another disilicide with a high thermoelectric power and has been proposed for application in a thermoelectric power generator. It has an orthorhombic crystal structure and is stable only at lower temperatures, whereas at higher temperatures it decomposes to form a two-phase FeSi – Fe_2Si_5 alloy. Furthermore, it is subject to pest oxidation. Doping with Al produces *p*-type conducting $FeSi_2$, doping with Co produces *n*-type $FeSi_2$, and the combination of both materials gives a thermocouple which can be used as a thermoelectric power generator. Such a power-generation system for the efficient conversion of low-grade heat into electricity has been proposed [12, 19].

The monosilicides MnSi and CoSi exhibit large thermoelectric powers and thus are promising for thermoelectric power generation. In particular, a CoSi – $CrSi_2$ thermogenerator has been proposed.

The complex manganese silicide $Mn_{11}Si_{19}$ is promising for thermoelectric applications because of its anomalously high anisotropy of the thermopower over an exceptionally wide temperature range [20].

7.2. Other Thermoelectric Compounds

Thermoelectric materials on the basis of the semiconducting compound Bi_2Te_3 are widely used as thermoelectric coolers (TECs) for operation near room temperature in small refrigerators [20] (→ Refrigeration Technology, Chap. 1.4). Such materials with the rhombohedral C33 structure are solid solutions of Bi_2Te_3 with the isomorphous compounds Sb_2Te_3 and Bi_2Se_3. They easily cleave along the basal planes and show anisotropic mechanical and transport properties.

Thermoelectric materials for application at medium temperatures are based on the compound PbTe with simple cubic B1 structure, which forms solid solutions with the isomorphous compounds PbSe, PbS, SnTe, and GeTe [20]. *n*- and *p*-type materials are prepared by deviation from stoichiometry and doping with halogens as donors and alkali metals as acceptors, respectively.

Research and development are leading to novel complex alloys with even more attractive properties [21, 22].

7.3. Thermomagnetic Compounds
[23 – 27]

Magnetization of a material by application of a strong magnetic field produces magnetic order with electron spin alignment and is accompanied by heat release. Without the magnetic field the material returns to magnetic disorder with heat consumption from the surroundings. This magnetocaloric effect is used for magnetic refrigeration, which is more efficient than conventional gas compression refrigeration. For this purpose magnetic refrigerant materials with high magnetic moments are needed, since the amount of produced or consumed heat is proportional to the magnetic moment apart from the magnetic field.

Intermetallic rare earth metal compounds with transition metals show very large magnetic moments and have successfully been used for magnetic refrigeration. Present developments concentrate on optimizing such rare earth metal phases with respect to magnetic moment and temperature of operation. An outstanding example is the phase $Gd_5Si_2Ge_2$ with a magnetocaloric effect which is twice as large as that of Gd [25]. The operating temperature of this material can be tuned from about −240 to about +20 °C by adjustment of the Si/Ge ratio.

The materials developments aim at magnetic refrigerators which can be applied, e.g., as small cryocoolers for cooling superconductive devices to near liquid helium temperatures or as large commercial liquefiers for liquefying cryofuels.

8. Optical Materials

Various intermetallics show beautiful colors which may be attractive for applications in jewelry. An early example is the Regulus of Venus of the alchemists with its bright violet color, which results from the reduction of Sb_2S_3, the

ore of antimony, through the addition of copper. A more recent example is the nickel disilicide $NiSi_2$, which can be alloyed with Ni and Al to produce compounds with distinct pale blue, yellow, and white colors and variations thereof. Au – Cu – Ag alloys on the basis of the intermetallic compounds CuAu and Cu_3Au, which have found applications in dentistry, excel by virtue of their decorative gold color. Likewise, PdIn with its gold color is promising for dental prosthetics and jewelry. A further example is the bright-purple $AuAl_2$ known as purple plague which gave rise to the Spangold alloy [28].

Apart from color, various intermetallics show optical properties such as reflectance, transmittance, absorbance and refraction, which are attractive for applications in optical devices [12]. GaAs is used for optical waveguides in integrated IR light circuits. Other applications of GaAs are optical high-power switches, high-efficiency solar cells, light-emitting diodes (LEDs), and semiconductor diode lasers. Various tellurides such as HgTe, CdTe, PbTe, SnTe, and corresponding solid solutions in which Zn and Mn beneficially substitute for Cd have been used for developing photodetectors for light from the far infrared to the visible region.

Thin films of various intermetallics such as $AuIn_2$, AuSn, InSb, BiTe, SnAs, CaSb, GeTe, and others are being developed for use as recording media in optical devices [12]. The local reflectance is changed by short laser beam pulses with subsequent slow or rapid cooling to produce a crystalline or amorphous structure, respectively. Likewise thin films of the shape-memory alloy NiTi (see Section 12.3) can be used as bistable optical memory elements.

Various ferromagnetic intermetallics are of interest in magneto-optical applications as materials for erasable optical data storage [29]. This relies on the polar Kerr effect, by which linearly polarized light experiences a rotation of the polarization plane when reflected at a magnetized material. The Kerr effect has been studied for a large number of Mn-based, Co-based, and Fe-based intermetallics. The highest potential for magneto-optical storage is attributed to MnBi and PtMnSb, besides amorphous TbFeCo-based alloys.

9. Hydrogen-Storage Materials

[30 – 35]

9.1. B2 Compounds

FeTi with B2 structure is one of the most promising intermetallic compounds for application as a hydrogen-storage material and related applications (see also → Hydrogen, Chap. 9.3.1.2). Dissolved hydrogen orders to form a substructure, whereby the B2 structure becomes rhombic, that is, a stable hydride is formed reversibly. FeTi with excess Ti is more easily hydrogenized then stoichiometric FeTi, since Ti activates the absorption and desorption of hydrogen catalytically. Second phases, including oxides, shorten the incubation time of hydrogenization, but oxygen may poison the FeTi surface at room temperature. Other B2 compounds of interest for applications as hydrogen-storage materials are CoTi and the ternary intermetallics (Fe,Co)Ti and (Fe,Ni)Ti, which show more advantageous storage properties than the binary compounds. The FeTi-base materials have good hydrogen capacities and low costs, but the problems with respect to activation, impurities, and instabilities have prevented large-scale commercial application.

9.2. Laves Phases

Various Laves phases can absorb large amounts of hydrogen and are considered for applications as hydrogen storage materials. Such AB_2 compounds contain a strong hydride former as the A element and crystallize with the C14 or C15 structure. The Laves phases with the highest sorption capacities are C15 ZrV_2 with a hydrogen-to-metal ratio H/M = 1.8, C14 or C15 $ZrCr_2$ with H/M = 1.3, C14 $ZrMn_2$ and C14 or C15 $TiCr_2$ with H/M = 1.2. $LaNi_2$, which was known as a C15 compound, with H/M = 1.5 is a tetragonal pseudo-Laves phase which is stabilized only by the presence of impurities according to recent findings.

The hydrides formed in these binary Laves phases are too stable for easy hydrogen desorption, which is a prerequisite for hydrogen storage. The hydride stability can be reduced and adjusted to practical hydrogen storage conditions by deviations from stoichiometry, substitution

of the B element, primarily by Fe, Co, Ni, Cu, Mn, Cr, substitution of the A element, primarily by Ti, or any combination of these alloying possibilities.

An example for the stoichiometry effect is given by $ZrMn_2$ which exhibits an increasing hydrogen vapor pressure with increasing Mn content and which is stable with C14 structure up to $ZrMn_{3.8}$. Another example is C14 $TiMn_2$, which reaches its maximum H solubility at the off-stoichiometric composition corresponding to $TiMn_{1.5}$, whereas it does not absorb H at the stoichiometric composition. The effects of substitution on electronic structure, alloy stability, hydride stability and H storage capacity have been studied for various systems, e.g., $(Zr_xTi_{1-x})Mn_2$, $Zr(Mn_{1-x}Fe_x)_{2+y}$, $Zr(V_{1-x}Co_x)_2$, in which both V and Co are substituted for by other transition metals, and $Zr(Cr_{1-x}V_x)_2$. $Zr(Cr_{1-x}V_x)_2$ exhibits polytypism, that is, crystal structure variants of either the hexagonal C14 structure or the cubic C15 structure are formed, which only differ in the stacking of crystal structure units. The thermodynamics of hydrogenation have been studied, and hydride formation enthalpies determined. The optimization of properties has led to multinary Laves phases such as Zr – Mn – Co – V alloys, Zr – Mn – Ni – V alloys, Zr – Mn – Ni – V – Cr alloys, or Zr – Ti – Mn – Fe and Zr – Ti – V – Fe alloys.

In summary, the Laves phase materials are advantageous because of their good hydrogen capacities in the temperature range 0 – 100 °C and low costs.

9.3. Rare Earth Metal Compounds and Other Phases

Various rare earth compounds can absorb large amounts of hydrogen and are of major interest for applications as hydrogen-storage materials, as is exemplified by RCo_3 and RFe_3. The most outstanding examples are the RNi_5 compounds, in particular $LaNi_5$, which may be regarded as a prototype hydrogen storage material (see also Chapter 10). The stability of $LaNi_5$ is affected by absorption – desorption cycling due to surface degradation processes. The property spectrum has been optimized by alloying with further elements, as is exemplified by the development of $La_{0.8}Nd_{0.2}Ni_{2.5}Co_{2.4}Si_{0.1}$ or $MmNi_{3.5}Co_{0.7}Al_{0.8}$ (Mm = mixed metal). The permanent-magnet compound $SmCo_5$ can also absorb large quantities of hydrogen which, however, reduces its coercive force. $CaNi_5$, which is isostructural with $LaNi_5$, shows an analogous behavior and is considered as a low-cost candidate for hydrogen storage.

Research and development aim at producing nanostructured materials, which can be obtained by powder-metallurgical processing through hydrogenation [32]. Such materials promise more advantageous storage properties because of their much-increased internal interface area and defect density, as demonstrated for Mg_2X phases (X = Cu, Ni, Si, Ge, Sn) [37, 38]. The prevailing problem is the hysteresis accompanying the hydriding and dehydriding processes, which reflects irreversibility and contributes to the degradation of the absorption properties during long-term service [31].

10. Electrode Materials

10.1. Mg_2Ni Alloys

Mg_2Ni-base hydrogen-storage alloys are of major interest for application as active materials of metal hydride electrodes. Such materials can be prepared by powder metallurgy, and their properties have been optimized by alloying with further elements [39]. Both partial substitutions of Ti or Ce for Mg and of Mn or Co for Ni improve discharge capacity and cycle life.

10.2. Laves Phase Alloys

Various rare earth (R) Laves phases undergo amorphization on hydrogenation. RNi_2 with C15 structure is easily transformed into amorphous RNi_2 which is very stable and is being considered for applications as long-life electrodes in electrochemical nickel – hydrogen batteries.

There are many other binary and ternary Laves phases which can absorb large amounts of hydrogen (see Section 9.2). The optimization of properties has led to multinary Laves phases such as Zr – Mn – Co – V alloys, Zr –

Mn – Ni – V alloys, Zr – Mn – Ni – V – Cr alloys, or Zr – Ti – Mn – Fe and Zr – Ti – V – Fe alloys, which have been studied in detail in view of applications as electrodes in electrochemical nickel – hydrogen batteries.

These developments have led to $Zr(Ni,Mn)_2$-base materials with high rate capability and high hydrogen capacity which are regarded as highly promising for application as negative electrode in nickel – metal hydride (NiMH) batteries for electric vehicles [36, 40].

10.3. RNi_5 Alloys

The rare earth (R) compounds RNi_5, in particular $LaNi_5$, of Section 9.3 are highly promising for applications as rechargeable electrodes in nickel – metal hydride (NiMH) batteries. The rare earths can be substituted by the cheaper Mm, and further alloying for optimiztion has led, for example, to $Mm(Ni,Co,Mn,Al)_5$, which is intended for application in high-power batteries [41]. Such batteries are presently commercialized and are used in newly developed pure electric vehicles (PEV) [33, 42 – 44].

Electrode materials are of particular interest for fuel cells, which are the subject of intense worldwide developments [45].

11. Coating Materials

Coatings are used in many cases to protect high-temperature materials in corrosive environments. Various coating systems have been developed which differ in processing and properties, as exemplified for high-temperature coatings for gas turbines [46, 47]. Such coatings usually contain intermetallic compounds as a major constituent. Present-day research and development is directed not only to high-temperature corrosion resistance, but also to high-temperature wear resistance and applications as thermal barriers [48 – 50].

11.1. Aluminides

An exceptionally high oxidation resistance is shown by the B2 compound NiAl. Accordingly NiAl has long been used as a coating material. Indeed, one of the oldest processes for producing coatings is pack aluminizing, in which Al diffuses from the Al-rich pack into the alloy to be coated and forms an aluminide surface layer, which consists of NiAl in the case of Ni-base alloys. Such diffusion coatings are still extensively used for protecting turbine blades under various turbine conditions. The coating performance is improved by adding alloying additions of Cr, Si, Ta, rare earths, or precious metals by modifying the pack aluminizing process.

Besides such diffusion coatings, overlay coatings are used which are produced by depositing a specifically designed corrosion-resistant alloy on the component surface (→ High-Temperature Materials). These coating alloys, which are described by the general formula MCrAlY (M = Fe, Ni, and/or Co), again rely on the presence of MAl phases with B2 structure. The advantage of these coatings is the negligibly small chemical interaction between coating and substrate during deposition, choice of corrosion-resistant alloy composition, and the ability to deposit thicker coatings for extended service lives. During service interdiffusion occurs between coating and substrate and may result in Al depletion of the coating. Al-deficient NiAl may transform martensitically because of cooling cycles, and this promotes spalling, that is, it contributes to the degradation of the coating.

Newly developed structural high-temperature materials may lack sufficient oxidation resistance, which initiates the development of protective coating systems. For example, the titanium aluminide alloys based on TiAl have insufficient oxidation resistance (see Section 16.1). Such alloys were coated successfully through application of a protective layer of the ternary intermetallic compound $Al_{22}Ti_8Cr_3$ with $L1_2$ structure by using the pack cementation technique [51]. Other successful coatings on TiAl alloys relied on applying a Ti – Al – Cr alloy for formation of the t phase together with the Laves phase [52] or an Ni – Al alloy with NiAl formation [53] as well as aluminizing with formation of the Al-rich titanium aluminide $TiAl_2$ [54].

11.2. Silicides

The alloys of the refractory metals molybdenum, niobium, and tantalum, which have major potential for structural applications at very high temperatures, must be protected from oxidation in oxidizing environments, and this is best achieved by applying silicide coatings [55]. Niobium alloys have successfully been coated by using a newly developed pack cementation process with codeposition of Si, Ti, Cr, and Fe, which relies on the formation of the quaternary silicide $Nb_3Fe_3CrSi_6$ with partial substitution of Nb by Ti and of Fe by Co or Ni [56].

Similar silicide diffusion coatings were developed for Cr – Nb alloys, which rely on the formation of $NbSi_2$ or $CrSi_2$, depending on doping with Ge, which suppresses formation of chromium disilicide [57]. Likewise a boron-modified titanium disilicide diffusion coating process with Ge doping by pack cementation was developed for protecting Ti alloys and Ti_3Al-based alloys (see Section 16.1) from oxidation [58]. TiAl-based alloys were successfully coated by codeposition of Al and Si by pack cementation [59] or simple liquid-phase siliconizing [60].

12. Shape-Memory Materials

[61 – 65]

12.1. Cu-Base Alloys

12.1.1. Cu – Zn – Al Shape Memory Alloys

The Cu – Zn – Al shape memory alloys are derived from the intermetallic compound CuZn with B2 structure by alloying with Al, and the composition of the resulting ternary compounds is approximately described by $Cu_{1+1.3x}Zn_{1-2.3x}Al_x$. The complete ternary Cu – Zn – Al phase diagram is available. The crystal structure of the Cu – Zn – Al alloy is the disordered A2 structure at high temperatures, which transforms into the B2 structure on cooling and into the $L2_1$ structure on further cooling. These structures are stable only at elevated temperatures and are only metastable at room temperature, that is, these alloys decompose if annealed at too low temperatures.

The $L2_1$ structure can be transformed martensitically to form various types of martensite. The crystallography and thermodynamic stability of the martensite variants have been analyzed in detail. The martensite stability is correlated with elastic anomalies, as in the case of CuZn. It was shown that different types of martensite can coexist for certain compositions. In any case, the martensite formation temperature is a very sensitive function of composition and can be varied by prior thermal cycling. In particular, alloying with Mn is used to depress the martensite formation temperature.

A special feature of the Cu – Zn – Al alloys is the so-called two-way memory effect which produces a reversible shape change on cooling and heating. This effect results from martensitic transformation and retransformation after specific thermomechanical treatments, which are designated as “training”.

Processing of Cu – Zn – Al alloys poses few problems. The alloys are usually produced by induction melting. Various dopants are in use for grain refinement to improve formability. Apart from this, powder metallurgy is used for obtaining fine-grained materials. The alloys are readily hot worked in air, and cold workability decreases with increasing Al content.

12.1.2. Cu – Al – Ni Shape-Memory Alloys

The Cu – Al – Ni shape memory alloys are based on the intermetallic compound Cu_3Al with disordered A2 structure, which is usually known as β-phase and which is stable only at high temperatures between 567 and 1049 °C with compositions between 71 and 82 atom % Cu. At 567 °C this phase decomposes by a eutectoid reaction at such a sluggish rate that it can be retained with the then metastable A2 structure by cooling below this temperature. At about 500 °C the β-phase undergoes an ordering reaction to form the metastable β_1-phase with $D0_3$ structure. Both phases can be transformed martensitically by quenching to form various types of martensite, depending on Al content and quench rate.

Cu_3Al can be alloyed with Ni without changing the transformation characteristics of the compound. Both the binary Cu_3Al and the ternary $(Cu,Ni)_3Al$ show the shape-memory

effect due to the formation of thermoelastic martensite. The ternary compound has given rise to the development of the Cu – Al – Ni shape memory alloys with about 11 – 15 wt % Al and 3 – 5 wt % Ni. The deformation modes of the martensites include various twinning modes and stress-induced transformations. The shape memory effect is a two-way effect, that is, it is completely reversible. The martensite-start temperature has an upper limit of about 180 °C and can be lowered to a large degree by increasing the Al content and to a smaller degree by increasing the Ni content.

The Cu – Al – Ni alloys are advantageous because of their higher stability at higher temperatures compared with the Cu – Zn – Al alloys. However, second-phase precipitation, which embrittles the Cu – Al – Ni alloys and precludes cold working, cannot be suppressed; hence, such alloys can only be hot finished. Thermomechanical treatments and microalloying additions, in particular Mn, Ti, Zr, are used for improving the mechanical behavior. An important effect is grain refinement, which increases both fracture strength and fracture strain according to a Hall – Petch relationship. Indeed precipitated particles have been found to reduce the as-cast grain size and hinder later grain growth during annealing. The alloying additions affect the transformation and decomposition behavior, and this must be taken into account in establishing the optimum conditions for the induction of the two-way effect. Recently, the ductilizing effect of small boron additions was discovered and studied in detail with respect to transformation behavior and microstructure.

Other similar Cu-base alloys also show the shape-memory effect. Well-studied examples are the Cu – Al – Mn shape memory alloys, which order by a two-step reaction on rapid cooling, that is, the initial disordered A2 structure is transformed first into the intermediate B2 structure and then into the final $L2_1$ structure.

12.2. NiAl Alloys

Al-deficient NiAl with B2 structure is transformed martensitically by rapid cooling. This transformation is thermoelastic and produces the shape-memory effect. The martensite formation temperature increases linearly with increasing Ni content from about − 240 °C for 60 atom % Ni to about 1000 °C for 70 atom % Ni. Thus, the shape-memory effect can be produced at high temperatures, and this allows the development of high-temperature shape-memory alloys. The martensitic transformation of Ni-rich NiAl can be induced by external applied stresses; it is thermoelastic, superelasticity has been observed, and the martensite deforms by twinning and detwinning. The transformation characteristics do not depend on the processing route. The problem of the low room-temperature ductility of NiAl has been overcome by alloying with third elements, in particular Fe, to produce a ductile second phase with fcc structure.

12.3. NiTi Alloys

The intermetallic compound NiTi with B2 structure has been used since about 30 years as shape-memory alloy for couplings, fasteners, connectors, and actuators in the automotive and aerospace industries, electronics, mechanical engineering, and medical applications. NiTi is known as Nitinol, which refers to Ni and Ti and the US Naval Ordnance Laboratory where this material was developed. NiTi melts congruently at 1310 °C and has an extended homogeneity range.

NiTi transforms martensitically from the parent B2 phase to a monoclinic martensite phase with an intermediate orthorhombic R phase, from which a reversible shape-memory effect with the best shape-memory behavior of all shape-memory alloys and pseudoelasticity results. The characteristics of such shape-memory alloys have been studied in detail with respect to crystallography, microstructure, thermodynamics, kinetics, and macroscopic mechanical behavior, and applications have been described.

The shape-memory behavior of NiTi is improved by prior thermomechanical treatments which produce advantageous microstructures and textures. The transformation temperature is about 110 °C for stoichiometric NiTi and is decreased by excess Ni. Likewise the transformation temperature is lowered by alloying with Fe and Cr, whereas Cu decreases the transformation temperature hysteresis. Recently a widened transformation temperature hysteresis

was found for Ni – Ti – Nb alloys which is attractive with respect to coupling and sealing applications. The transformation temperature is increased to 500 °C by the partial substitution of Ni by Pd, which, however, affects ductility.

The elastic and plastic deformation behavior, including mechanical twinning, of NiTi alloys has been studied intensively. Noteworthy is the high ductility of NiTi, which allows deformations of up to 70 %. Furthermore the yield strength of NiTi depends on temperature in an anomalous way, that is, it increases with increasing temperature to reach a maximum, and only above this peak temperature does normal softening occur.

Processing of NiTi alloys is difficult and must be done with care to obtain the advantageous properties. Melting must be carried out in vacuum or under inert atmosphere because of the reactivity of Ti, and hence vacuum-induction, plasma-arc, and electron-beam melting are all used commercially. Hot working is possible without problems. Processing at low temperatures is difficult because of extremely high work hardening, and intermediate annealing treatments are required for cold rolling, machining, etc. The ductility problem of the (Ni,Pd)Ti alloys with high transition temperature is relieved by alloying with B. Joining by welding, brazing, or soldering is generally difficult. The excellent shape-memory behavior of the NiTi alloys is accompanied by an excellent corrosion resistance, comparable to stainless steels, which makes them the only shape-memory alloys suitable for implantation in the human body. Recently NiTi has been used as strengthening phase in Al matrix composites, which experience an additional strengthening effect due to the shape-memory effect of NiTi.

13. Medical Biomaterials [66 – 69]

Classical metallic biomaterials for medical implants and other devices are stainless steels, in particular the austenitic AISI 316L steel, Co – Cr(– Mo) alloys including Vitallium, manifold Ti alloys, i.e., commercially pure (cp) Ti, Ti6Al4V and other Ti alloys with Zr and/or Nb and/or Ta, cp Nb, cp Ta [70], and there are developments for using Mg alloys [71]. Such alloys are used usually for applications such as orthopedic implants, nails, screws, plates for fixing bones, artificial hip joints, heart valve prostheses, and stents. Problems are posed by the necessary biocompatibility, which is controlled primarily by corrosion resistance and physical alloy properties. The corrosion resistance increases with the transition from steels to Co – Cr alloys and further to Ti alloys which, however, is paralleled by respective cost increases.

In view of the advantageous corrosion behavior of the Ti alloys, intermetallic Ti – Al alloys (see Section 16.1.2) have recently attracted major interest for endoprosthetic applications, such as hip and knee prostheses and dental implants [72].

Besides the presently used biomaterials, the intermetallic phase NiTi known as Nitinol is extensively used for a wide variety of applications because of its outstanding attractive properties [73 – 75]. In addition to being biocompatible, this material is superelastic and shows the shape-memory effect (see Section 12.3). The extremely good biocompatibility is due to the formation of a passive TiO2 layer. Since NiTi is not ferromagnetic, NiTi devices are compatible with magnetic resonance imaging (MRI).

Superelastic NiTi wires are used for connecting bone fragments and fixing small bone fragments. Other applications are filters for large embolized blood clots. The superelasticity gives rise to an enormous flexibility which is made use of for producing medical instruments, e.g., superelastic NiTi catheters. The most celebrated application is the self-expanding stent, which relies on knitted or welded NiTi wire. The long-term stability of such devices is still the subject of discussion, and in particular there is the question whether toxicity may arise after many years of service.

14. Dental Materials

14.1. Amalgams

Dental amalgams have been in use for more than a thousand years. Amalgams are still a most important class of material for dental restoration in spite of corrosion problems with mercury release because they offer an advantageous combination of properties with respect to processing and use.

These amalgams are formed by alloying Ag – Sn – Cu alloys on the basis of the intermetallic compounds Ag_3Sn and Cu_3Sn with mercury. The amalgamation reaction results in the formation of a complex multiphase alloy structure which contains the phases Ag_2Hg_3, Sn_8Hg, Ag_3Sn, and Cu_6Sn_5, depending on composition.

Processing involves mixing and compaction of powders and is comparatively simple and is analogous to mechanical alloying. The amalgams must show a sufficient compressive strength and creep resistance, since the oral temperature is only about 10 % below the melting temperature. The dental amalgams are very brittle which leads to failure with tensile or bending stresses in unsupported margins. The fracture processes are correlated with corrosion processes which primarily attack the phases Sn_8Hg and Cu_6Sn_5 and lead to the release of mercury and SnO, respectively. Attempts to replace mercury in these amalgams have not yet been successful.

The alloy systems of interest have been studied in detail with respect to phase reactions, microstructure evolution, mechanical behavior including creep and fatigue, corrosion resistance and clinical performance, and the processing technologies are well established [76, 77].

14.2. Cu – Au Alloys

Au – Cu – Ag alloys on the basis of the intermetallic compounds CuAu and Cu_3Au have found applications in dentistry because of their extremely high corrosion resistance, their advantageous mechanical properties such as high strength and ductility, and their decorative gold color. These alloys age-harden as a result of complex ordering and decomposition reactions by which the phases Cu_3Au I, CuAu I, CuAu II, and an Ag-rich α_2 phase are formed, depending on composition.

Cu_3Al I is the well-known Cu_3Al compound with $L1_2$ structure, which is the prototype phase for this crystal structure. Cu_3Au II forms at intermediate temperatures with off-stoichiometric compositions and is described as either tetragonal or orthorhombic. In any case, the Cu_3Au II structure results from the $L1_2$ structure by the introduction of equally spaced antiphase boundaries (APBs). Such a structure is usually known as long-period superlattice (LPS). At the equiatomic composition the tetragonal $L1_0$ phase CuAu I forms, and again an orthorhombic LPS is observed at intermediate temperatures, which is known as CuAu II. All phases are separated from each other by two-phase equilibria.

The elastic behavior is known, as is the diffusion behavior. The deformation behavior of Cu_3Au was studied in detail to clarify the effects of atomic order, and it was found in particular that order decreases the yield stress and increases the work hardening rate while still retaining high ductility. Inversely, heavy cold deformation destroys the order. The ordering reaction produces antiphase domains which give rise to an additional strengthening effect in this and similar phases; strength increases with decreasing domain size. The domain size can be varied by prior processing, and it increases during ageing in analogy to Ostwald ripening. The Portevin – Le Chatelier effect with serrated yielding was observed for both the ordered and disordered state. Recovery and recrystallization have been analyzed in detail.

15. Wear-Resistant Materials

Intermetallic compounds are attractive in tribological applications because of various advantageous properties, such as hardness, thermal stability, and oxidation resistance, and because of their compatibility with metallic matrices [78]. Some of the oldest bearing alloys contain intermetallic compounds as is exemplified by the lead-based bearing alloys (babbits) with SbSn and possibly Cu_6Sn_5 or the bronze bearing alloys with $Cu_{32}Sn_8$ or Cu_3Al.

The ternary Laves phase MoCoSi, that is, $Mo(Co,Si)_2$, with hexagonal C14 structure has given rise to the development of wear-resistant Co – Mo – Cr – Si alloys which contain large volume fractions of the Laves phase in coarse distribution together with a Co-rich phase and are known as Tribaloys. Apart from their high strength and hardness, they show a sufficient fracture toughness of the order of 20 $MN/m^{3/2}$. These advantageous mechanical properties result in an excellent wear resistance which, in combination with an excellent corrosion resistance in various environments, makes the Tribaloys very appealing for applications in a broad

range of temperatures and environments. These alloys have been studied in detail including corrosive wear testing, according to which the Tribaloys compare favorably with conventional wear-resistant materials [78]. The low ductility of Tribaloys, which is due to the large amounts of Laves phase, can be increased by proper adjustment of alloy composition and heat treatment [79, 80].

Apart from the Tribaloys, modified structural intermetallic high-temperature alloys (see Chap. 16.1) have been proposed for application as wear-resistant alloys. An example is the novel Nicralc alloys which result from alloying Ni3Al with C to increase the hardness and which are intended as substitutes for high-temperature abrasion- and erosion-resistant cobalt alloys [81]. Other examples are superelastic NiTi (see Section 12.3) with hardening TiC dispersoids [82] and MoSi2 – SiC [83] and Fe_3Al – $Fe_3AlC_{0.5}$ composites [84].

16. Structural High-Temperature Materials

The aluminides of titanium, iron, and nickel, as well as some silicides, have been selected as bases for developments of new structural high-temperature alloys because of their high strengths at low and high temperatures, high thermal stability, high oxidation resistance, and/or low density. The problem of insufficient ductility was overcome by further alloying and optimization of microstructure, which usually resulted in multiphase alloys [85]. Research and development is in progress, and some developments are on the brink of commercialization. The basic properties of the intermetallic compounds in question and the respective phase diagrams, as well as the physical, mechanical and chemical properties of the respective developed or proposed intermetallic alloys, with emphasis on powder metallurgy processing, are summarized in [4].

16.1. Titanium Aluminide Alloys

16.1.1. Ti_3Al Alloys

The titanium aluminide Ti_3Al, often designated as the α_2 phase, shows the hexagonal $D0_{19}$ structure and a density of 4.2 g/cm^3, which make it most attractive for materials developments (→ Titanium, Titanium Alloys, and Titanium Compounds, Chap. 11.2).

The aim of the various materials developments on the basis of Ti_3Al has been to improve both strength and ductility by alloying with further elements and by controlling the microstructure. The most effective element for improving ductility is Nb, and indeed Ti_3Al-base alloys with engineering significance, that is, α_2 alloys and super-α_2 alloys, contain 10 – 30 atom % Nb. Ti_3Al-base alloys readily dissolve oxygen and hydrogen, which both produce embrittlement.

The developments of the Ti_3Al-based alloys are on the brink of commercialization. They have been produced in ingots of up to 4500 kg and are available as cast shapes or wrought forms including billets, bars, plates, and sheet. Likewise the powder metallurgy methods which have been developed for conventional titanium alloys are readily adapted and in addition other powder metallurgy methods, in particular reactive processing, have been applied to Ti_3Al-base alloys. The alloys can be machined, sheet can be deformed superplastically and diffusion bonded to form complex sheet parts, and the various manufacturing processes can be accomplished on conventional equipment. The alloys can be joined successfully by both diffusion bonding and linear friction welding.

16.1.2. TiAl Alloys

The titanium aluminide TiAl, often designated the γ-phase, crystallizes with the tetragonal $L1_0$ structure. The density of 3.76 g/cm^3 is even lower than that of Ti_3Al and has made the titanium aluminides most attractive for materials developments (→ Titanium, Titanium Alloys, and Titanium Compounds, Chap. 11.2). Single-crystalline TiAl shows an anomalous positive temperature dependence of the yield stress, that is, the yield stress increases with increasing temperature until a maximum is reached at about 600 °C, and only above the peak temperature is a normal negative temperature dependence observed.

The major problem for the application of TiAl is low ductility at room temperature. If the Al content is reduced below 50 atom % to form Ti_3Al as second phase besides TiAl, the resulting intermetallic two-phase alloys show ductilities of a few percent which are acceptable for applications. 48 atom % Al is the optimum composition for maximum room-temperature ductility of TiAl alloys. The aim of the various materials developments on the basis of TiAl is to improve both strength and ductility first by controlling the microstructure and second by alloying with further elements. Oxygen embrittles TiAl alloys, as do other interstitial impurities (N, C, B). In addition, environmental embrittlement has been found for TiAl, as for other intermetallics, which is clearly caused by the presence of atmospheric moisture.

The development of TiAl alloys is less advanced than that of the Ti_3Al-based materials because of the more severe ductility problems. However, the TiAl alloys are regarded as more promising, primarily because of the lower density and the higher application temperatures, and development activities are now concentrated on TiAl in various countries. Processing is more difficult compared with Ti_3Al alloys because of the higher brittle-to-ductile transition temperature, and hot working temperatures are outside the range of conventional titanium processing equipment, which has limited the practical ingot size to about 200 kg. Besides casting and hot working, powder metallurgy has been used for component fabrication, though there are problems of quality, availability, and cost of powder.

Cast TiAl alloys can be forged isothermally, fine-grain powder metallurgy material can be formed superplastically, as well as special cast and thermomechanically treated material, sheet of 1 – 2 mm thickness can be produced by twin-roll casting, special processes have been developed for rolling, and machining is possible, as is joining by diffusion bonding and welding. With respect to turbine-engine applications, component-fabrication techniques have been demonstrated for airfoils and compressor cases. Besides aerospace applications, interest is focused on automotive engine applications such as valves [86, 87] and turbocharger rotors [88, 89], which are nearer to realization.

More recent materials developments concentrated on Ti – Al alloys with reduced Al contents and significantly increased Nb contents of about 5 – 10 atom % for relieving the ductility problems [90].

16.2. Iron Aluminide Alloys

16.2.1. Fe_3Al Alloys

Fe_3Al is formed on cooling by ordering reactions in the solid state that transform the bcc disordered solid solution, which is stable above about 800 °C, first into the FeAl phase with B2 structure, which is stable between about 800 °C and 550 °C, and only then into Fe_3Al with the $D0_3$ structure. The critical temperature of ordering is much lower than the melting temperature, and this indicates weaker bonding between unlike atoms than in, for example, the nickel aluminides. The critical temperature of $D0_3$ ordering can be shifted to higher temperatures by as much as 250 °C by alloying with third elements, in particular Cr, Mo, Mn, Ti, and Si. Alloying with Ti is of particular interest in view of the high solubility of Ti in Fe_3Al with beneficial effects on strength and ductility [91]. Thermal vacancies are formed readily because of a low formation enthalpy, which is even lower than the migration enthalpy, and consequently there is a high concentration of vacancies at thermal equilibrium. High concentrations of vacancies formed at high temperatures are trapped even by slow cooling and contribute to strengthening. Strength and creep resistance at high temperatures can be increased significantly by proper alloying with precipitation of second phases, as has been exemplified by Fe_3Al with precipitated perovskite-type carbide and/or Laves phase particles [92, 93].

Fe_3Al is considered for structural applications since it not only shows a higher strength than comparable iron alloys, but also a high corrosion resistance in oxidizing and sulfidizing environments. Problems for application are the low ductility at ambient temperatures and the strength drop above the ordering temperature. The ductility of Fe_3Al is very sensitive to the environment, and environmental embrittlement can occur in the presence of moisture in the ambient atmosphere. Alloys have been produced by both powder metallurgy and ingot metallurgy,

including arc melting, vacuum-arc remelting, induction melting in air and vacuum, and electroslag remelting. These developments, aimed at applications such as automotive exhaust systems and resistance heating elements, are close to commercialization. Other potential applications are hot-gas filters in coal gasification plants and chemical processing equipment.

16.2.2. FeAl Alloys

FeAl is closely related to NiAl (Section 16.3.2) since both phases show the B2 structure and a complete mutual miscibility. However, FeAl melts incongruently in contrast to NiAl, and its melting temperature is lower, which indicates a lower stability and atomic bonding strength. The tensile ductility at low temperatures of polycrystalline FeAl is practically zero at the stoichiometric composition and increases with decreasing Al content, that is, increasing deviation from stoichiometry. The reported ductilities of off-stoichiometric FeAl are obtained only with well-annealed specimens, that is, if the low-temperature deformation is preceded by annealing at very low cooling rates. This behavior is supposed to be due to retained excess vacancies, which increase hardness and yield strength and decrease ductility and which are not equilibrated after high-temperature annealing with normal cooling rates. Anneals of several hundred hours at about 400 °C are recommended for producing ductility in off-stoichiometric FeAl. FeAl is subject to environmental embrittlement, which is due to the dissolution of hydrogen in FeAl at the crack tip and is reduced by increasing the deformation rate sufficiently.

In view of its reported advantageous properties and comparatively low density, FeAl is regarded as a potential structural material for high-temperature applications, and processing developments make use of both ingot and powder metallurgy. FeAl alloy compositions have been identified, e.g., for good weldability and for high corrosion and wear resistance in aggressive gasification environments. As in the case of Fe_3Al, the high-temperature strength and creep resistance of FeAl can be improved significantly by proper alloying with precipitation of hardening second phases [94, 95].

16.3. Nickel Aluminide Alloys

16.3.1. Ni_3Al alloys

Ni_3Al is the most studied and best known intermetallic compound because it has long been used as strengthening phase in the superalloys and because its ductility problems were overcome by ductilization by microalloying with boron. Hence, it can be produced and tested without major experimental difficulties, and thus it was preferred in the past for studying the behavior of intermetallics. Ni_3Al, known as the γ'-phase, crystallizes with the cubic $L1_2$ structure.

The outstanding feature of the plastic flow of Ni_3Al is the anomalous temperature dependence of the flow stress, which increases between room temperature and about 700 °C with increasing temperature to reach a maximum, and only at higher temperatures does normal softening occur. Polycrystalline Ni_3Al is brittle and fails by intergranular fracture, whereas Ni_3Al single crystals are highly ductile. The polycrystal brittleness of Ni_3Al is removed by microalloying Al-deficient Ni_3Al with B.

Exposure to air or oxygen affects the mechanical behavior of Ni_3Al, and various effects of environmental embrittlement have been observed. At about 800 °C the ductility of ductilized Ni_3Al shows a minimum which is more pronounced in air than in vacuum and which is significantly alleviated by the addition of Cr. Environmental embrittlement has also been observed at room temperature, especially in moist air, depending on Ni_3Al composition and dopant content, and it was concluded that this embrittlement is caused by hydrogen formation. Accordingly, stress corrosion cracking has been observed in aqueous environments.

Successful materials developments on the basis of Ni_3Al have resulted in a series of closely related Ni_3Al alloys for high-temperature applications. These alloys, known as nickel aluminides or advanced aluminides, generally contain B at levels below 500 ppm and Al below the stoichiometric composition to give ductility at room temperature, as well as up to 5 atom % Hf, Zr, Ta, and Mo for improving strength at high temperatures and up to 10 atom % Cr for

enhancing ductility at intermediate temperatures between 400 °C and 900 °C.

The Ni_3Al alloys compare favorably with many superalloys with respect to mechanical properties. Thus, Ni_3Al alloys enlarge the superalloy spectrum. However, they cannot compete with the advanced aerospace superalloys used in aircraft engines. Nevertheless, Ni_3Al alloys are promising for less demanding applications where good combinations of strength and resistance against fatigue, wear, including erosion and cavitation, and oxidation are needed. Ni_3Al alloys can be melted by air-induction melting with acceptable quality and by vacuum-induction melting, vacuum-arc remelting, and electroslag remelting for better qualities. Ingots up to 1500 kg and components have been cast with low or no porosity, depending on boundary conditions, hot working is possible by conventional hot forging, isothermal forging, and hot extrusion, the alloys can be cold worked, and property data are available. Most of the alloys can be welded with good quality, provided welding is performed with care and cutting is accomplished by high-speed abrasive wheels. In addition, powder-metallurgy methods have been applied successfully. The transfer process from the laboratory to large-scale production is handicapped by the limited deformability at elevated temperatures, which makes hot working difficult. The feasibility of commercial large-scale manufacturing has successfully been demonstrated by producing rolls of a Ni_3Al alloy for a steel hardening furnace by conventional foundry practice and automated welding [96].

16.3.2. NiAl Alloys

NiAl is the best known example of the intermetallic compounds with the cubic B2 structure which form one of the largest group of intermetallics. NiAl has an extended range of homogeneity and melts congruently at about 1640 °C for the stoichiometric composition with 50 atom % Al. This melting point is higher than those of the constituent elements, and this indicates strong bonding between Ni and Al and correspondingly a high phase stability with a strong tendency for atomic ordering.

NiAl shows high strength at low temperatures, a comparatively steep strength decrease at intermediate temperatures around half the melting temperature of about 680 °C for NiAl, and low strength with further softening at high temperatures. The softening at high temperatures is related to thermally activated creep processes. Deviations from stoichiometry produce constitutional defects which enhance diffusion and lead to softening at high temperatures, whereas at low temperatures these defects act as strengthening deformation obstacles. Polycrystalline NiAl is usually brittle with practically no ductility.

Unlike most other aluminides, NiAl with B2 structure exhibits excellent oxidation resistance since a protective Al_2O_3 scale is readily formed during oxidation. NiAl seems to be the only truly oxidation resistant intermetallic besides some silicides. The physical reason for this high oxidation resistance is that the Al content is sufficiently high and Al diffusion is sufficiently fast in NiAl at all temperatures for stable Al_2O_3 scales to form at the surface and to avoid internal oxidation in the bulk. Environmental embrittlement, which affects the mechanical behavior of various other intermetallics (e.g., the other B2 aluminide FeAl) has not been found for NiAl.

NiAl is advantageous for high-temperature applications because of its low density, good thermal conductivity, high melting point, and excellent oxidation resistance, whereas its limited deformability at room temperature and its low strength and creep resistance at temperatures above 1000 °C are disadvantages compared to the superalloys. The low strength at high temperatures is related to the excellent oxidation resistance, since both effects result from easy high-temperature diffusion. NiAl is regarded as highly promising for developing alloys which are competitive with superalloys, because the strength deficit is compensated by advantages in density and thermal conductivity, and respective materials developments are under way.

As conventional disordered alloys, NiAl alloys can be strengthened appreciably by second phases. Apart from the effects on strength, second phases may also be beneficial for ductility and toughness. An example for the latter is the successful development of an NiAl – Fe – Cr alloy with sufficient ductility at room temperature and strength at low and high temperatures for applications in industrial furnaces, coal-

conversion systems, and petrochemical plants [97].

For the strengthening of NiAl second phases with not too low Al contents are preferred with respect to density and oxidation resistance. There are hard and brittle ternary Laves phases with Al, such as NbNiAl and TaNiAl, which may be used for strengthening NiAl. A respective development has lead to a NiAl – Ta – Cr alloy which is not deformable at room temperature but excels in high-temperature strength, oxidation resistance, and shock resistance [98]. The alloy can be produced by both ingot metallurgy and powder metallurgy, so that components can be produced by investment casting, hot extrusion, hot isostatic pressing (HIP), powder injection molding, and isothermal forging.

16.4. Silicide Alloys

Silicides (→ Silicon, Chap. 5) were considered for high-temperature applications as early as 50 years ago because of their high strength and oxidation resistance at high temperatures. Presently the need for new structural materials for the highest temperatures has revived interest in silicides for structural applications, and respective developments are under way.

Cr_3Si with the topologically close packed cubic A15 structure is a candidate phase for high-temperature applications. The brittle Cr_3Si is alloyed with Cr to produce a ductile-phase-toughened intermetallic composite [99]. Ni_3Si with the ordered fcc $L1_2$ structure has a very high potential for structural applications because of its advantageous mechanical properties and its outstanding corrosion resistance. Alloying of Ni_3Si with Nb and Ti results in multiphase Ni_3Si alloys with dispersed Ni_3Nb and Ni_3Ti precipitates which are most promising for high-temperature applications because of their favorable mechanical and chemical properties [100].

Mg_2Si crystallizes with the fcc C1 structure, and its density is only 1.88 g/cm^3. Mg_2Si has a comparatively high strength and a low thermal expansion coefficient. However, its brittleness, with a brittle-to-ductile transition temperature of 450 °C, precludes the application of single-phase Mg_2Si. Alloying with Al produces Al-Mg_2Si alloys with brittle Mg_2Si particles in a ductile Al matrix, which has given rise to a successful alloy development.

The M_5Si_3 compounds (M = transition metal), which adopt a variety of complex crystal structures, are usually very stable congruently melting line compounds with melting temperatures above 2000 °C. The most attractive M_5Si_3 compound for structural high-temperature application is Ti_5Si_3 with the hexagonal $D8_8$ structure because of its high stability with a melting temperature of 2130 °C, its high strength and hardness, and its low density. A promising materials development is based on the combination of Ti_3Al and Ti_5Si_3 which has been further alloyed with Nb. Another candidate phase for high-temperature applications is Nb_5Si_3, which is even more stable with a melting temperature of 2484 °C and which is alloyed with Nb to relieve its brittleness. Such Nb – Nb_5Si_3 composites are alloyed with further elements to achieve an excellent balance of room-temperature toughness, high-temperature creep performance, and oxidation resistance [101, 102].

Presently, $MoSi_2$ is considered for structural high-temperature applications, and various materials developments are in progress. The brittle-to-ductile transition occurs at about 1000 °C or higher temperatures depending on microstructure and impurity content. The ductility and toughness of $MoSi_2$ can be improved by special processing routes, microalloying with C for avoiding SiO_2 layers on grain boundaries or insertion of a ductile second phase, usually an Nb-rich phase. Toughness can also be increased by embedding hard second phases, in particular ceramics, in a $MoSi_2$ matrix to hinder crack growth. $MoSi_2$-based composites reinforced by TiC, TiCN, and TiB_2 indeed showed improved hardness and bending strength compared to monolithic $MoSi_2$ [103]. Most work has concentrated on the $MoSi_2$ – SiC system. Other systems of particular interest are $MoSi_2$ – TiC and $MoSi_2$ – Al_2O_3, as well as the eutectics $MoSi_2$ – Mo_5Si_3 and $MoSi_2$ – $Er_2Mo_3Si_4$ and the ductile-phase-toughened $MoSi_2$ – Nb/Ta and $MoSi_2$ – Mo systems. More recent developments concentrate on intermetallic Mo – Si – B alloys consisting primarily of the phases Mo_3Si and Mo_5SiB_2 besides ductile-phase-toughening Mo [104 – 106]. Such composites show attractive oxidation resistances, fracture toughness, and creep strength, the bal-

ance of which depends sensitively on composition adjustment and processing. Furthermore, $MoSi_2$ has been considered as reinforcing phase in other composite systems, e.g. in an SiC matrix or in a beryllide matrix. $TiSi_2$ with orthorhombic C54 structure has also been considered for structural applications at high temperatures because of its high oxidation resistance and low density.

17. References

1. G. E. R. Schulze: *Metallphysik*, Akademie-Verlag, Berlin 1967.
2. G. Sauthoff: *Intermetallics*, VCH Verlagsgesellschaft, Weinheim 1995.
3. N. S. Stoloff, V. K. Sikka (eds.): *Physical Metallurgy and Processing of Intermetallic Compounds*, Chapman & Hall, London 1996.
4. G. Sauthoff: "Intermetallic Materials", in P. Beiss, R. Ruthardt, H. Warlimont (eds.): *Landolt-Börnstein (New Series) Group VIII Subvolume VIII/2A - Powder Metallurgy Data*, Part 2, Refractory Hard and Intermetallic Materials, Springer, Berlin 2001, pp. 221–267.
5. J. H. Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds, 4 Volume Set*, John Wiley & Sons, Chichester 2000.
6. J. H. Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds, Principles and Practice, Vol. 3, Progress*, John Wiley & Sons, Chichester 2002.
7. K. J. Hemker et al. (eds.): *Structural Intermetallics 2001 (Proc. ISSI-3)*, TMS, Warrendale 2001.
8. J. H. Schneibel, K. J. Hemker, R. D. Noebe, S. Hanada, G. Sauthoff (eds.): *High-Temperature Ordered Intermetallic Alloys IX*, MRS, Warrendale 2001.
9. P. Villars: *Pearsons's Handbook Desk Edition — Crystallographic Data for Intermetallic Phases*, ASM International, Materials Park/OH 1997.
10. K. H. J. Buschow: *Permanent Magnet Materials and their Applications*, TTP Transtech Publications, Ütikon-Zürich 1998.
11. H. Harada et al.: "Magnetic Materials," in P. Beiss, R. Ruthardt, H. Warlimont (eds.): *Landolt-Börnstein (New Series) Group VIII: Advanced Materials and Technologies, Subvolume VIII/2A: Powder Metallurgy Data, Part 1: Metals and Magnets*, Springer-Verlag, Berlin 2003, pp. 9.1–9.83.
12. J. H. Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds, Vol. 4, Magnetic, Electrical and Optical Properties and Applications of Intermetallic Compounds*, John Wiley & Sons, Chichester 2000.
13. R. Harris: "The Attractions of Rare-Earth Magnets," *Materials World* **7** (1999) 478 – 481.
14. K. Tachikawa, A. Nagata, K. Togano: "Advances in Superconducting Materials," *J. Jpn. Inst. Metal.* **66** (2002) 199–200.
15. K. S. Kumar: "Silicides: Science, Technology and Applications", in J. H. Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds: Principles and Practice*, Vol. 2, Wiley, Chichester 1995, pp. 211 – 236
16. S. P. Murarka: "Silicide Thin Films and Their Applications in Microelectronics," *Intermetallics* **3** (1995) 173–186.
17. J. H. Westbrook: "Miscellaneous Applications", in J. H. Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds: Principles and Practice*, Vol. 2, Wiley, Chichester 1995, pp. 645 – 656.
18. K. Masumoto, A. Katsui, T. Matsuoka: "Semiconductor Applications", in J. H. Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds: Principles and Practice*, Vol. **2**, Wiley, Chichester 1995, pp. 323 – 350.
19. K. Ono, R. O. Suzuki: "Thermoelectric-Power Generation — Converting Low-Grade Heat into Electricity," *JOM J. Min. Met. Mat. Soc.* **50** (1998) 49 – 51.
20. M. V. Vedernikov: "Thermoelectric and Electrical Applications", in J. H. Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds: Principles and Practice*, Vol. 2, Wiley, Chichester 1995, pp. 453 – 472.
21. G. Nicolas, J. Sharp, J. Goldsmid: *Thermoelectrics*, Springer, Berlin 2001.
22. G. Chen, M. S. Dresselhaus, G. Dresselhaus, J. P. Fleurial, T. Caillat: "Recent Developments in Thermoelectric Materials," *Int. Mater. Rev.* **48** (2003) 45–66.
23. J. A. Barclay: "Active and Passive Magnetic Regenerators in Gas Magnetic Refrigerators," *J. Alloys Compounds* **207** (1994) 355–361.
24. E. Bruck: "Developments in Magnetocaloric Refrigeration," *J. Phy. D Appl. Phys.* **38** (2005) R381–R391.
25. K. A. Gschneidner, V. K. Pecharsky, A. O. Tsokol: "Recent Developments in Magnetocaloric Materials," *Rep. Progr. Phys.* **68** (2005) 1479–1539

26. M. Ilyn et al.: "Magnetocaloric Properties of the LaFe11.7Si1.3 and LaFe11.2Co0.7Si1.1 Systems," *J. Magn. Magn. Mater.* **290** (2005) 712–714.
27. D. Vasylyev, J. Prokleska, J. Sebek, V. Sechovsky: "Thermal Properties of Er(Co1-xSix)(2) Compounds," *J. Alloys Compounds* **394** (2005) 96–100.
28. I. M. Wolff, M. B. Cortie: "The Development of Spangold," *Gold Bull.* **27** (1994) 44–54.
29. W. A. McGahan: "Magneto-Optical Applications", in J. H. Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds: Principles and Practice*, Vol. 2, Wiley, Chichester 1995, pp. 435 – 452.
30. V. A. Yartys, I. R. Harris, V. V. Panasyuk: "New Metal-Hydrides — A Survey," *Mater. Sci. (Engl. Transl.)* **37** (2001) 219–240.
31. P. Dantzer: "Properties of Intermetallic Compounds Suitable for Hydrogen Storage Applications," *Mater. Sci. Eng. A Struct. Mater.* **329** (2002) 313–320.
32. M. Okada, T. Kuriiwa, A. Kamegawa, H. Takamura: "Role of Intermetallics in Hydrogen Storage Materials," *Mater. Sci. Eng. A Struct. Mater.* **329** (2002) 305–312.
33. R. A. Varin, T. Czujko: "Overview of Processing of Nanocrystalline Hydrogen Storage Intermetallics by Mechanical Alloying/Milling," *Materials and Manufacturing Process* **17** (2002) 129–156.
34. A. Zuttel: "Materials for Hydrogen Storage," *Mater. Today* **6** (2003) 24–33.
35. M. Conte, P. P. Prosini, S. Passerini: "Overview of Energy/Hydrogen Storage: State-of-the-Art of the Technologies and Prospects for Nanomaterials," *Mater. Sci. Eng. B* **108** (2004) 2–8.
36. D. M. Kim, K. J. Jang, J. Y. Lee: "A Review on the Development of AB(2)-Type Zr-Based Laves Phase Hydrogen Storage Alloys for Ni-MH Rechargeable Batteries in the Korea Advanced Institute of Science and Technology," *J. Alloys Compd.* **293 – 295** (1999) 583–592.
37. H. Y. Shao, Y. T. Wang, H. R. Xu, X. G. Li: "Preparation and Hydrogen Storage Properties of Nanostructured Mg2Cu Alloy," *J. Solid State Chem.* **178** (2005) 2211–2217.
38. R. Janot, F. Cuevas, M. Latroche, A. Percheron-Guegan: "Influence of Crystallinity on the Structural and Hydrogenation Properties of Mg2X Phases (X=Ni, Si, Ge, Sn)," *Intermetallics* **14** (2006) 163–169.
39. J. Chen, P. Yao, D. H. Bradhurst, S. X. Dou, H. K. Liu: "Mg_2Ni-based Hydrogen Storage Alloys for Metal Hydride Electrodes," *J. Alloys Compd.* **293 – 295** (1999) 675 – 679.
40. B. Knosp, L. Vallet, P. Blanchard: "Performance of an AB_2 Alloy in Sealed Ni-MH Batteries for Electric Vehicles — Quantification of Corrosion Rate and Consequences on the Battery Performance," *J. Alloys Compd.* **293–295** (1999) 770 – 774.
41. A. Otto, V. Güther: "Development of Fast Kinetics Metal Hydride Alloys and Battery Electrodes for High Power Applications," *J. Alloys Compd.* **293 – 295** (1999) 734 – 736.
42. T. Sakai, I. Uehara, H. Ishikawa: "Research-and-Development on Metal Hydride Materials and Ni-MH Batteries in Japan," *J. Alloys Compd.* **293 – 295** (1999) 762 – 769.
43. F. Zhan, L. J. Jiang, B. R. Wu, Z. H. Xia, X. Y. Wei, G. R. Qin: "Characteristics of Ni/MH Power Batteries and its Application to Electric Vehicles," *J. Alloys Compd.* **293 – 295** (1999) 804 – 808.
44. J. M. Joubert, M. Latroche, A. Percheron-Guegan: "Metallic Hydrides II: Materials for Electrochemical Storage," *MRS Bull.* **27** (2002) 694–698.
45. G. Marsh: "Fuel Cell Materials," *Mater. Today* **4** (2001) 20–24.
46. P. K. Datta, J. S. Burnell-Gray, K. Natesan: "Coating Technology," in J. H Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds — Principles and Practice, Vol. 3 Progress*, John Wiley & Sons, Chichester 2002, pp. 561–588.
47. J. R. Nicholls, R. Wing: "Advances in Coating Systems for Utility Gas Turbines," in J. Lecomte-Beckers, M. Carton, F. Schubert, P. J. Ennis (eds.): Materials for Advanced Power Engineering 2002 (Proc. 7th Liège Conference), Forschungszentrum Jülich, Jülich, 2002, pp. I.57–I.71.
48. A. Matthews, J. H. Givens, W. L. Ensinger, S. L. Rohde (eds.): *Metallurgical Coatings and Thin Films 1999 (Proc. 26th Int. Conf. Metallurgical Coatings and Thin Films)*, North-Holland Elsevier Science, Amsterdam 2000.
49. J. R. Nicholls: "Designing Oxidation-Resistant Coatings," *JOM J. Min. Met. Mat.* **52** (2000) 28–35.
50. A. G. Evans, D. R. Mumm, J. W. Hutchinson, G. H. Meier, F. S. Pettit: "Mechanisms Controlling the Durability of Thermal Barrier

Coatings," *Prog. Mater. Sci.* **46** (2001) 505–553.
51. H. Mabuchi, H. Tsuda, T. Kawakami, S. Nakamatsu, T. Matsui, K. Morii: "Oxidation-Resistant Coating for Gamma-Titanium Aluminides by Pack Cementation," *Scr. Mater.* **41** (1999) 511 – 516.
52. G. S. Fox-Rabinovich, D. S. Wilkinson, S. C. Veldhuis, G. K. Dosbaeva, G. C. Weatherly: "Oxidation Resistant Ti-Al-Cr Alloy for Protective Coating Applications," *Intermetallics* **14** (2006) 189–197.
53. T. Izumi, T. Nishimoto, T. Narita: "Superior Long-Term Oxidation Resistance of Ni-Al Coated TiAl Alloys," *Intermetallics* **13** (2005) 727–732.
54. Z. Liu, G. Wang: "Improvement of Oxidation Resistance of [gamma]-TiAl At 800 and 900 [deg]C in Air by TiAl2 Coatings," *Mate. Sci. Eng. A* **397** (2005) 50–57.
55. C. M. Packer: "Overview of Silicide Coatings for Refractory Metals", in T. Grobstein, J. Doychak (eds.): *Oxidation of High-Temperature Intermetallics*, TMS, Warrendale 1989, pp. 235 – 244.
56. M. Vilasi, M. Francois, R. Podor, J. Steinmetz: "New Silicides for New Niobium Protective Coatings," *J. Alloys Compd.* **264** (1998) 244 – 251.
57. Y. R. He, R. A. Rapp, P. F. Tortorelli: "Oxidation-Resistant Ge-Doped Silicide Coating on Cr-Cr_2Nb Alloys by Pack Cementation," *Mater. Sci. Eng. A* **222** (1997) 109 – 117.
58. B. Cockeram, R. A. Rapp: "Development and Growth of Boron Modified and Germanium Doped Titanium Silicide Diffusion Coatings by the Halide Activated, Pack Cementation Method," *Oxid. Met.* **45** (1996) 375–425.
59. Z. D. Xiang, S. R. Rose, P. K. Datta: "Codeposition of Al and Si to Form Oxidation-Resistant Coatings on [gamma]-TiAl by the Pack Cementation Process," *Mater. Chem. Phys.* **80** (2003) 482–489.
60. H. P. Xiong, W. Mao, Y. H. Xie, Y. Y. Cheng, X. H. Li: "Formation of Silicide Coatings on the Surface of a TiAl-Based Alloy and Improvement in Oxidation Resistance," *Mater. Sci. Eng. A* **391** (2005) 10–18.
61. J. Vanhumbeeck: "High-Temperature Shape-Memory Alloys," *J. Eng. Mater. Technol. Trans. ASME* **121** (1999) 98 – 101.
62. J. Vanhumbeeck: "Nonmedical Applications of Shape-Memory Alloys," *Mater. Sci. Eng. A Struct. Mater.* **275** (1999) 134–148.
63. R. D. James, K. F. Hane: "Martensitic Transformations and Shape-Memory Materials," *Acta Mater.* **48** (2000) 197–222.
64. J. Vanhumbeeck: "Shape-Memory Alloys — A Material and a Technology," *Adv. Eng. Mater.* **3** (2001) 837–850.
65. D. S. Grummon: "Thin-Film Shape-Memory Materials for High-Temperature Applications," *JOM J. Min. Met. Mat.* **55** (2003) 24–32.
66. J. A. Helsen, H. J. Breme (eds.): *Metals as Biomaterials*, John Wiley & Sons Ltd, Chichester, New York 1998.
67. A. Fischer: "Biomaterials — Fundamentals and Clinical-Applications," *Materialwiss. Werkstofftech.* **32** (2001) 109.
68. J. B. Park, J. D. Bronzino (eds.): *Biomaterials: Principles and Applications*, CRC Press, Boca Raton, FL, 2002.
69. B. D. Ratner, A. Hoffman, F. Schoen, J. Lemons (eds.): *Biomaterials Science: An Introduction to Materials in Medicine*, Academic Press, New York 2004.
70. K. S. Katti: "Biomaterials in Total Joint Replacement," *Colloids Surf. B: Biointerf.* **39** (2004) 133–142.
71. M. P. Staiger, A. M. Pietak, J. Huadmai, G. Dias: "Magnesium and its Alloys as Orthopedic Biomaterials: A Review," *Biomaterials* **27** (2006) 1728–1734.
72. M. L. Escudero, M. A. Munoz-Morris, M. C. Garcia-Alonso, E. Fernandez-Escalante: "In Vitro Evaluation of a [gamma]-TiAl Intermetallic for Potential Endoprothesic Applications," *Intermetallics* **12** (2004) 253–260.
73. T. Duerig, A. Pelton, D. Stockel: "An Overview of Nitinol Medical Applications," *Mater. Sci. Eng. A Struct. Mater.* **275** (1999) 149–160.
74. F. Elfeninat, G. Laroche, M. Fiset, D. Mantovani: "Shape-Memory Materials for Biomedical Applications," *Adv. Eng. Mater.* **4** (2002) 91–104.
75. A. R. Pelton, S. M. Russell, J. DiCello: "The Physical Metallurgy of Nitinol for Medical Applications," *JOM J. Min. Met. Mat.* **55** (2003) 33–37.
76. R. M. Waterstrat, T. Okabe: "Dental Amalgams", in J. H. Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds: Principles and Practice*, Vol. 2, Wiley, Chichester 1995, pp. 575 – 590.

77. C. Guminski: "Amalgams," in J. H. Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds — Principles and Practice, Vol. 3 Progress*, John Wiley & Sons, Chichester 2002, pp. 21–35.
78. W. A. Glaeser: "Intermetallics in Tribology", in J. H. Westbrook, R. L. Fleischer (eds.): *Intermetallic Compounds: Principles and Practice*, Vol. 2, Wiley, Chichester 1995, pp. 591 – 602.
79. R. Liu, W. Xu, M. X. Yao, P. C. Patnaik, X. J. Wu: "A Newly Developed Tribaloy Alloy with Increased Ductility," *Scripta Mater.* **53** (2005) 1351–1355.
80. R. Liu, M. X. Yao, P. C. Patnaik, X. J. Wu: "Effects of Heat Treatment on Mechanical and Tribological Properties of Cobalt-Base Tribaloy Alloys," *J. Mater. Eng. Perform.* **14** (2005) 634–640.
81. H. Goldenstein, Y. Nunes Silva, H. N. Yoshimura: "Designing a New Family of High Temperature Wear Resistant Alloys Based on Ni3Al IC: Experimental Results and Thermodynamic Modelling," *Intermetallics* **12** (2004) 963–968.
82. D. Y. Li: "Development of Novel Tribo Composites with TiNi Shape Memory Alloy Matrix," *Wear* **255** (2003) 617–628.
83. P. La, Q. Xue, W. Liu: "Study of Wear Resistant MoSi2-SiC Composites Fabricated by Self-Propagating High Temperature Synthesis Casting," *Intermetallics* **11** (2003) 541–550.
84. J. Yang, P. La, W. Liu, J. Ma, Q. Xue: "Tribological Properties of Fe3Al-Fe3AlC0.5 Composites under Dry Sliding," *Intermetallics* **13** (2005) 1184–1189.
85. G. Sauthoff: "Multiphase Intermetallic Alloys for Structural Applications," *Intermetallics* **8** (2000) 1101 – 1109.
86. M. Blum, G. Jarczyk, H. Scholz, S. Pleier, P. Busse, H. J. Laudenberg, K. Segtrop, R. Simon: "Prototype Plant for the Economical Mass Production of TiAl-Valves," *Mater. Sci. Eng. A Struct. Mater.* **329** (2002) 616–620.
87. K. Gebauer: "Performance, Tolerance and Cost of TiAl Passenger Car Valves," *Intermetallics* **14** (2006) 355–360.
88. M. Nazmy, V. Lupinc: "Gamma TiAl Intermetallics for Turbomachinery Applications," in J. Lecomte-Beckers, M. Carton, F. Schubert, P. J. Ennis (eds.): Materials for Advanced Power Engineering 2002 (Proc. 7th Liège Conference), Forschungszentrum Jülich, Jülich, 2002, pp. I.43–I.56.
89. T. Tetsui: "Development of a TiAl Turbocharger for Passenger Vehicles," *Mater. Sci. Eng. A Struct. Mater.* **329** (2002) 582–588.
90. F. Appel: "Recent Developments in the Design and Processing of Gamma-Based Titanium Aluminide Alloys," in 4th International Conference on Processing and Manufacturing of Advanced Materials, JUL 07-11, 2003 (THERMEC'2003), PTS 1-5, 2003, pp. 91–98.
91. M. Palm, G. Sauthoff: "Deformation Behaviour and Oxidation Resistance of Single-Phase and Two-Phase L21 Fe-Al-Ti Alloys," *Intermetallics* **12** (2004) 1345–1359.
92. D. G. Morris, M. A. Munoz-Morris, C. Baudin: "The High-Temperature Strength of Some Fe3Al Alloys," *Acta Mater.* **52** (2004) 2827–2836.
93. M. Palm, A. Schneider, F. Stein, G. Sauthoff: "Strengthening of Iron Aluminide Alloys for High-Temperature Applications," in M. J. Mills, H. Inui, H. Clemens, C. L. Fu (eds.): *Integrative and Interdisciplinary Aspects of Intermetallics*, MRS, Warrendale, 2005, pp. 3–14.
94. D. G. Morris, M. A. Munoz-Morris, J. Chao: "Development of High Strength, High Ductility and High Creep Resistant Iron Aluminide," *Intermetallics* **12** (2004) 821–826.
95. D. D. Risanti, G. Sauthoff: "Iron-Aluminium-Base Alloys with Strengthening Laves Phase for Structural Applications at High Temperatures," *Mater. Sci. Forum* 475–479 (2005) 865–868.
96. V. K. Sikka et al.: "Large-Scale Manufacturing of Nickel Aluminide Transfer Rolls for Steel Austenitizing Furnaces," *Intermetallics* **12** (2004) 837–844.
97. D. Letzig, J. Klöwer, G. Sauthoff: "Screening of NiAl-Base Ni-Fe-Al Alloys for Structural High Temperature Applications and Development of a New Ni-30Fe-10Al-Cr Alloy," *Z. Metallkd.* **90** (1999) 712 – 721.
98. M. Palm, G. Sauthoff: "Characterization and Processing of an Advanced Intermetallic NiAl-Base Intermetallic Alloy for High-Temperature Applications," in K. J. Hemker et al. (eds.): *Structural Intermetallics 2001 (Proc. ISSI-3)*, TMS, Warrendale, 2001, pp. 149–156.
99. H. Bei, G. M. Pharr, E. P. George: "A Review of Directionally Solidified Intermetallic Composites for High-Temperature Structural

Applications," *J. Mater. Sci.* **39** (2004) 3975–3984.
100. K. Ohira, Y. Kaneno, T. Takasugi: "Microstructure, Mechanical Property and Oxidation Property in Ni3Si-Ni3Ti-Ni3Nb Multi-phase Intermetallic Alloys," *Mater. Sci. Eng. A* **399** (2005) 332–343.
101. B. P. Bewlay, M. R. Jackson, J. C. Zhao, P. R. Subramanian: "A Review of Very-High-Temperature Nb-Silicide-Based Composites," *Metall. Mater. Trans. A* **34** (2003) 2043–2052.
102. C. L. Ma, J. G. Li, Y. Tan, R. Tanaka, S. Hanada: "Microstructure and Mechanical Properties of Nb/Nb5Si3 In Situ Composites in Nb-Mo-Si and Nb-W-Si Systems," *Mater. Sci. Eng. A* **386** (2004) 375–383.
103. J. Meng, J. Lu, J. Wang, S. Yang: "Preparation and Properties of MoSi2 Composites Reinforced by TiC, TiCN, and TiB2," *Mater. Sci. Eng. A* **396** (2005) 277–284.
104. S. Nakatani, K. Yoshimi, S. Hanada, S. H. Ko, Y. H. Park: "Design of Mo-Si-B-O Pseudo in-situ Composites for High Temperature Applications," in PRICM 4: Fourth Pacific Rim International Conference on Advanced Materials and Processing, Vols I and II, 2001, pp. 971–974.
105. J. H. Schneibel et al.: "Optimization of Mo-Si-B intermetallics," in E. P. George, H. Inui, M. J. Mills, G. Eggeler (eds.): *Defect Properties and Related Phenomena in Intermetallic*, MRS, Warrendale 2003, pp. 53–58.
106. V. Behrani, A. J. Thom, M. J. Kramer, M. Akinc: "Microstructure and Oxidation Behavior of Nb-Mo-Si-B Alloys," *Intermetallics* **14** (2006) 24–32.

Mössbauer Spectroscopy

PHILIPP GÜTLICH, Institut für Anorganische Chemie und Analytische Chemie, Johannes Gutenberg-Universität, Mainz, Federal Republic of Germany

JÜRGEN ENSLING, Institut für Anorganische Chemie und Analytische Chemie, Johannes Gutenberg-Universität, Mainz, Federal Republic of Germany

1. Introduction 1
2. Principle and Experimental Conditions of Recoil-free Nuclear Resonance Fluorescence 1
2.1. Nuclear Resonance 1
2.2. Recoil Effect 2
2.3. Mössbauer Effect 3
3. Mössbauer Experiment 5
3.1. Doppler Effect 5
3.2. Mössbauer Spectrometer 5
4. Preparation of Mössbauer Source and Absorber 7
5. Hyperfine Interactions 8
5.1. Isomer Shift 8
5.2. Magnetic Hyperfine Interaction 10
5.3. Quadrupole Splitting 11
5.4. Combined Quadrupole and Magnetic Hyperfine Interactions 13
5.5. Relative Intensities of Resonance Lines 14
6. Evaluation of Mössbauer Spectra . . . 14
7. Selected Applications 14
7.1. Magnetic Materials 14
7.2. Corrosion Products 16
7.3. Catalysis 16
7.4. Atmospheric Aerosol Samples 16
7.5. U.S. One-Dollar Bill 17
8. References 17

1. Introduction

Since the discovery of the Mössbauer effect by RUDOLF MÖSSBAUER [1–3] in 1958 this nuclear spectroscopic method has found a wide variety of applications in materials science, solid state physics, chemistry, metallurgy, and earth sciences.

So far, the Mössbauer effect has been observed for almost 100 nuclear transitions in more than 40 different elements. Not all of these transitions are suitable for actual studies, but ca. 15 – 20 elements remain that are appropriate for applications, of which the best known Mössbauer nuclide is iron, because the Mössbauer experiment is easy to perform with the 14.4 keV transition of ^{57}Fe, and iron is very abundant in nature.

This article introduces the principles of Mössbauer spectroscopy and the various types of chemical information that can be extracted from the electric and magnetic hyperfine interactions reflected in the Mössbauer spectrum. The introductory part focuses on the principles of ^{57}Fe Mössbauer spectroscopy, and then examples of typical applications are presented. Further information can be found in [4–13].

2. Principle and Experimental Conditions of Recoil-free Nuclear Resonance Fluorescence

2.1. Nuclear Resonance

The resonance absorption of electromagnetic radiation is a phenomenon well known in many branches of physics. The scattering of sodium light by sodium vapor, the excitation of a tuning fork by sound, and of a dipole by radiofrequency radiation are some familiar examples. The possibility of observing the phenomenon in nuclear transitions involving γ radiation was suggested in 1929 by KUHN [14]. However, for many years attempts to detect it remained unsuccessful, until RUDOLF MÖSSBAUER established the experimental conditions as well as the theoretical description of the occurrence of nuclear resonant absorption (fluorescence) of γ-rays [1–3]. The following conditions must be fulfilled for the observation of the Mössbauer effect.

Consider a nucleus in an excited state of energy E_e with the mean lifetime τ. Transition to the ground state of energy E_g occurs by emitting a γ quantum with the energy

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.e16_e01

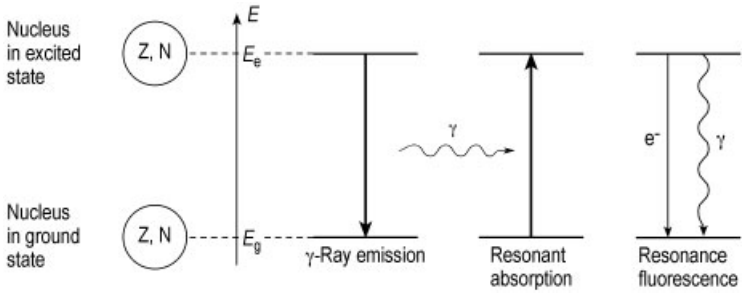

Figure 1. Schematic representation of nuclear resonance absorption of γ-rays (Mössbauer effect) and nuclear resonance fluorescence

$$E_0 = \hbar w = E_e - E_g \tag{1}$$

This γ quantum may be absorbed by a nucleus of the same kind in its ground state, which thereby takes up the energy difference E_0. This phenomenon is called nuclear resonant absorption of γ-rays and is schematically depicted in Figure 1.

The fact that the lifetime τ of an excited nuclear state is finite means that its energy has a certain distribution of width Γ, which, due to the Heisenberg uncertainty relationship, is connected to τ by

$$\tau\Gamma = \hbar \tag{2}$$

Consequently, the probability for the emission of a quantum is a function of the energy, or, in other words, the energy spectrum exhibits an intensity distribution $I(E)$ about a maximum at E_0 with a width at half maximum Γ which corresponds to the uncertainty in energy of the excited nuclear state. The explicit shape of this distribution is Lorentzian according to

$$I(E) \propto \frac{\Gamma/2\pi}{(E - E_0)^2 + (\Gamma/2\pi)^2} \tag{3}$$

The mean lifetime τ of the excited states of the Mössbauer isotopes are in the range of 10^{-6} to 10^{-10} s; the corresponding uncertainties of the transition energies are between 10^{-9} and 10^{-5} eV. Nuclear resonance fluorescence has a remarkably high energy resolution, given by the ratio Γ/E_0, which is on the order of 10^{-12}. As an example, the first excited state of ^{57}Fe has a mean lifetime $\tau = 141$ ns, which leads to a natural linewidth $\Gamma = \hbar/\tau = 4.7 \times 10^{-9}$ eV and hence to a relative energy resolution of $\Gamma/E_0 = 3.3 \times 10^{-13}$.

The absorption line is also described by a Lorentzian curve. Nuclear resonant absorption can only be observed when the difference of the transition energies for emission and absorption are not much greater than the sum of the linewidths, i.e., when the emission and absorption lines overlap sufficiently.

After resonance absorption, the excited nucleus will decay by either emitting isotropically a γ quantum (as in the primary γ-ray emission of Figure 1) or a conversion electron e^-, preferentially from the K shell. This phenomenon is termed *nuclear resonance fluorescence* and may be used in Mössbauer scattering experiments (surface investigations).

2.2. Recoil Effect

Consider the emission of a γ quantum with energy E_γ from an excited nucleus of mass M in a free atom or molecule (in the gas or liquid phase). The energy of the excited state is E_0 above that of the ground state. If the excited nucleus is assumed to be at rest before the decay, a recoil effect will be imparted to the nucleus upon γ emission, causing the nucleus to move with velocity v in opposite direction to that of the emitted γ-ray (see Fig. 2).

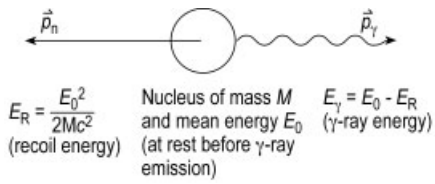

Figure 2. Recoil of momemtum $\vec{p}_n$ and energy E_R imparted to an isolated nucleus upon γ-ray emission

The resulting recoil energy is

$$E_{\mathrm{R}} = \frac{1}{2}Mv^2 = \frac{p_{\mathrm{n}}^2}{2M} \tag{4}$$

Momentum conservation requires that

$$p_\gamma = E_\gamma/c = -p_{\mathrm{n}} \tag{5}$$

where p_γ and p_{n} are the linear momenta of the γ quantum and the recoiled nucleus, respectively, and c is the velocity of light. Hence, the energy of the emitted γ quantum is

$$E_\gamma = E_0 - E_{\mathrm{R}} \approx E_0 \tag{6}$$

Since E_{R} is very small compared to E_0, we may write for the recoil energy (in a nonrelativistic approximation)

$$E_{\mathrm{R}} = \frac{p_{\mathrm{n}}^2}{2M} = \frac{E_\gamma^2}{2Mc^2} \approx \frac{E_0^2}{2Mc^2} \tag{7}$$

For many nuclear transitions $E_{\mathrm{R}} \approx 10^{-2}$–$10^{-3}$ eV. The recoil effect causes the energy of the emitted photon to be decreased by E_{R}, and the emission line appears at an energy $E_0 - E_{\mathrm{R}}$ rather than at E_0. In the absorption process, the γ-ray to be absorbed by a nucleus requires the total energy $E_0 + E_{\mathrm{R}}$ to make up for the transition from the ground to the excited state and the recoil effect, which causes the absorbing nucleus to move in the same direction as the incoming γ-ray photon. As shown schematically in Figure 3 the transition lines for emission and absorption are separated by $2E_{\mathrm{R}}$, which is about 10^6 times larger than the natural line width Γ.

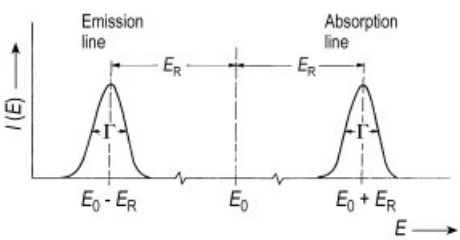

Figure 3. Consequences of the recoil effect caused by γ-ray emission and absorption in isolated nuclei. The transition lines for emission are separated by $2E_{\mathrm{R}} \approx 10^6\,\Gamma$. There is no overlap between emission and absorption lines, and hence no resonant absorption is possible

The condition for overlap of emission and absorption line, i.e., $2E_{\mathrm{R}} \leq \Gamma$, is not fulfilled. Thus, the cross section of nuclear resonant absorption in isolated atoms and molecules in the gaseous and liquid phases is extremely small.

In optical spectroscopy, however, the recoil energy is negligible compared with the linewidth, and the resonance condition is certainly fulfilled for freely emitting and absorbing particles.

Before the discovery of the Mössbauer effect some successful experiments on nuclear resonance absorption in the gas phase had been carried out utilizing the Doppler effect. High velocities were required ($\approx 10^2$ – 10^3 m s^{-1}). This could be achieved either by mounting a source of emitting nuclei on a high speed rotor or by heating the source and the absorber, so that some overlap of the lines could take place.

Rudolf Mössbauer, while working on his Ph. D. Thesis, carried out experiments of this kind with ^{191}Ir. But surprisingly, on lowering the temperature he found an increase in the absorption effect rather than a decrease. Mössbauer was able to explain this unexpected phenomenon, and was awarded the Nobel Prize in 1961 for the observation and correct interpretation of the recoilless nuclear resonance absorption (fluorescence) which is now known as the *Mössbauer effect* and which provided the basis for a powerful technique in solid state research.

2.3. Mössbauer Effect

As explained in Section 2.2, nuclear resonance absorption and fluorescence of γ-rays is not possible for nuclei in freely emitting or absorbing particles. This is different in the solid state. The nuclei are more or less rigidly bound to the lattice. If a γ-ray photon is emitted from the excited nucleus, the concomitant recoil energy may now be assumed to consist of two parts,

$$E_{\mathrm{R}} = E_{\mathrm{tr}} + E_{\mathrm{vib}} \tag{8}$$

E_{tr} refers to the translational energy transferred through linear momentum to the crystallite of mass M. As M is about 10^{20} times larger than the mass of a single atom, the corresponding recoil energy E_{tr} imparted to the crystallite becomes negligible in comparison with γ (Eq. 7).

Most of the recoil energy E_{R} will be dissipated into processes other than linear momentum, i.e., to the lattice vibrational system. E_{vib},

the vibrational energy part, still being of the order of 10^{-3} eV and thus several orders of magnitude smaller than the atom displacement energy (ca. 25 eV) and also much smaller than the characteristic lattice phonon energies (ca. $10^{-1} - 10^{-2}$ eV), is dissipated by heating the nearby lattice surroundings. As a consequence of this phonon creation with frequencies up to w_i, the resulting energy of the emitted γ-ray is

$$E_\gamma = E_0 - \sum_i n\hbar w_i \tag{9}$$

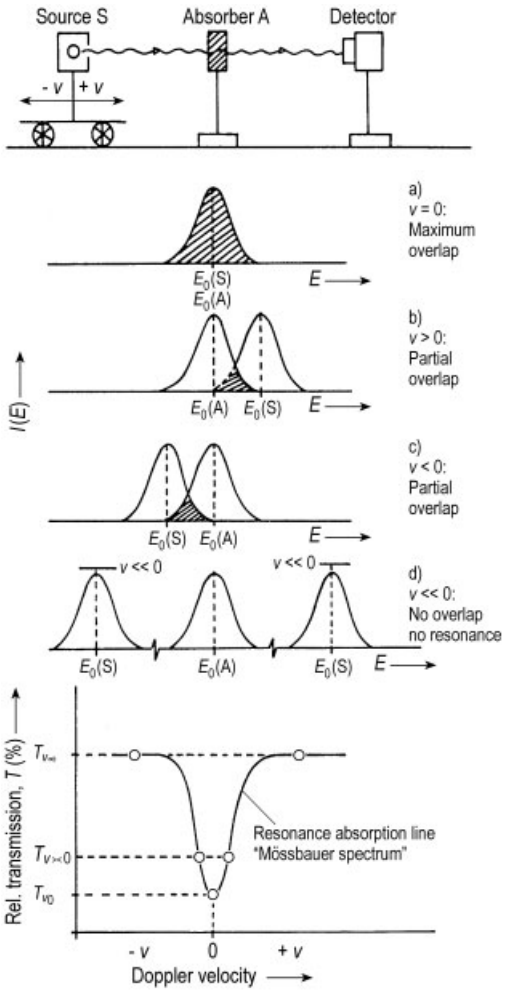

Figure 4. Schematic illustration of the experimental arrangement for recoilless nuclear resonance absorption, and relative transmission of recoilless γ quanta as a function of Doppler velocity

This again would destroy the resonance phenomenon by shifting the emission and absorption lines too far apart from each other on the energy scale. Fortunately, however, the lattice vibrations are quantized, and E_{vib} changes the vibrational energy of the lattice oscillators by integral multiples of the phonon energy, $n\hbar w_i$ ($n = 0, 1, 2, \ldots$). This means that there is a certain probability f that no lattice excitation (energy transfer of $0 \cdot \hbar w_i$, called the zero-phonon process) occurs during the γ-ray emission or absorption process. Mössbauer showed that only to the extent of this probability f of unchanged lattice vibrational states during γ-ray emission and absorption does nuclear resonance absorption become possible. This is termed the *Mössbauer effect*. The probability f is called the recoil-free fraction and denotes the fraction of nuclear transitions which occur without recoil.

The recoil-free fraction in Mössbauer spectroscopy is equivalent to the fraction of X-ray scattering processes without lattice excitation; this fraction of elastic processes in X-ray and neutron diffraction is described by the Debye – Waller factor:

$$f = e^{-k^2 \langle x^2 \rangle} \tag{10}$$

where $\langle x^2 \rangle$ is the mean square amplitude of vibration of the resonant nucleus in the photon propagation direction and $k = 2\pi/\lambda = E_\gamma/(\hbar c)$ is the wavenumber of the γ quantum. The value of f depends on the lattice properties, described by $\langle x^2 \rangle$, and on the photon properties, represented by k^2. The ratio of the mean square displacement and wavelength λ of the photon must be small ($\ll 1$), in order to have a sufficiently high Debye – Waller factor f. Clearly, the Mössbauer effect cannot take place in liquids, where the molecular motion is characterized by unbound $\langle x^2 \rangle$, causing the recoil-free fraction to vanish. The following consequences emerge:

1) f decreases exponentially with k^2, i.e., E_γ^2. The Mössbauer effect can hardly be detected with γ-ray energies above 150 keV.
2) The mean square displacement $\langle x^2 \rangle$ should be small; in other words, the Mössbauer atom should be bound tightly to the crystal lattice. However, crystallinity of the material is not a necessary condition for observation of the Mössbauer effect. The effect is also detectable in amorphous materials, glasses, and in frozen solutions.

3) The recoil-free fraction f is temperature dependent, because $\langle x^2 \rangle$ decreases with decreasing temperature and hence forces f to increase. However, f does not reach unity even at $T = 0$, since $\langle x^2 \rangle \neq 0$ at $T = 0$ because of the quantum-mechanical zero-point motion of the atoms (nuclei). To express f in terms of the usual experimental variables, the mean square amplitude $\langle x^2 \rangle$ is calculated by using lattice dynamic models (e.g., Einstein, Debye) [6], [15].
4) $\langle x^2 \rangle$ and hence f can be anisotropic. As a consequence, a dependence of the intensity of absorption lines on the observation direction in single-crystal experiment may be observed.

3. Mössbauer Experiment

3.1. Doppler Effect

The main components of a Mössbauer spectrometer are the γ-ray source containing the Mössbauer-active nuclides, the absorber (or scatterer), and a detector for low-energy γ radiation (see Fig. 4) plus electronics for automatic recording of the spectrum. The source and the absorber are moved relative to each other (either by moving the source and keeping the absorber fixed or vice versa). The transmitted (scattered) γ quanta are recorded as a function of the relative velocity. In this way, it is possible to trace, stepwise, the absorption line by the emission line utilizing the Doppler effect. A γ quantum which is emitted from the source moving at a velocity v receives a Doppler energy E_{D} modulation such that

$$E_\gamma = E_0 + E_{\mathrm{D}} = E_0 \left(1 + \frac{v}{c}\right) \tag{11}$$

where E_0 is the energy of the γ quantum emitted from the same source at rest and $c$is the velocity of light. This relationship is valid for a moving source, the velocity being positive for the source moving towards the absorber, and negative for moving away from the absorber. Owing to the very narrow linewidth Γ of the resonance line, it is generally sufficient to produce a small Doppler energy change E_{D} of the order of the line width Γ by Doppler shifting the source in order to sweep over the resonance. The Doppler velocities needed in the case of ^{57}Fe spectroscopy are in the range 0 to $\pm\, 10\,\mathrm{mm\,s^{-1}}$, and for most of the other Mössbauer nuclides velocities of less than $\pm\, 100\,\mathrm{mm\,s^{-1}}$ are sufficient.

The Mössbauer spectrum, a plot of the relative transmission as a function of Doppler velocity, shows maximum resonance and therefore minimum relative transmission at relative velocities where emission and absorption lines overlap ideally (Fig. 4). At high positive or negative velocities, the overlap of emission and absorption lines is negligible, the resonance effect being practically zero, i.e., the relative transmission yields the base line.

3.2. Mössbauer Spectrometer

Modern Mössbauer spectrometers generate the spectrum by the velocity-sweep method. The drive system moves the source (or absorber) repeatedly over a range of velocities, while simultaneously counting the γ quanta behind the absorber into synchronized channels.

The essential components of a modern Mössbauer spectrometer, as illustrated in the block-diagram of Figure 5, are the velocity transducer, the wave form generator and synchronizer, the multichannel analyzer, γ-ray detection system, a cryostat or oven for low-temperature and temperature-dependent measurements, a velocity calibration device, the source, the absorber, and a read-out unit.

The source (and in a few cases the absorber) is mounted on the vibrating axis of an electromagnetic transducer (loudspeaker system) [16] which is moved according to a voltage waveform applied to the driving coil of the system. The usual velocity functions are of the triangular, sawtooth, or sinusoidal form. In special cases the source may be moved with a constant velocity.

A function generator produces the desired waveform of the motion. A feedback system (feedback amplifier) operates in such a manner that the actual velocity of the source follows the reference waveform of the function generator with an accuracy better than 0.1 %. Using the triangular waveform one obtains an undistorted spectrum, because the velocity is a linear function of time. In extreme cases, where high velocities are required or the mass to be moved

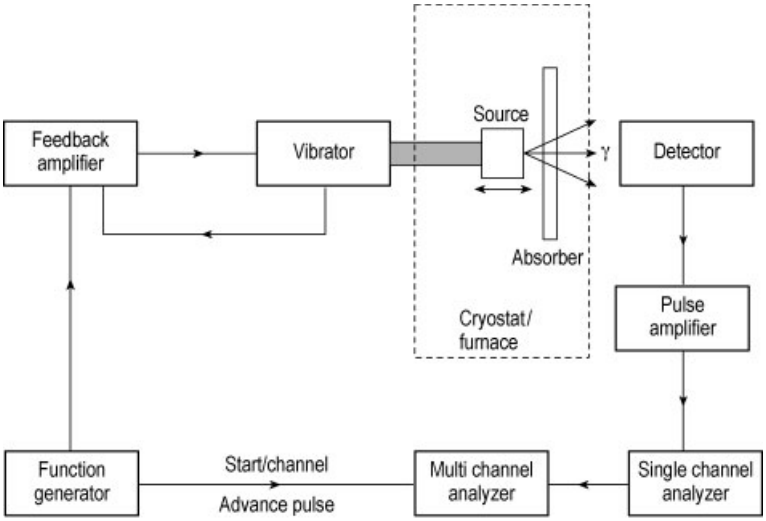

Figure 5. Block diagram illustrating the principle of a Mössbauer spectrometer

is large, it is preferable to use a sinusoidal waveform. For convenience, this type of spectrum is converted after the measurement into the linear form.

The counts detected behind the absorber are stored in a memory (mostly in the memory of a mutlichannel analyzer containing typically 512 or 1024 channels). Synchronization of the channel number in the memory and the instantaneous velocity of the source (absorber) is achieved by advancing the address of the memory one by one through an external clock which subdivides the period of the waveform into the number of available channels. A start pulse, which coincides with the beginning of the waveform, triggers the multichannel analyzer to start with channel no. 1, which is subsequently advanced by the clock pulses. In this way synchronization is achieved, i.e., during each period of motion a certain channel of the analyzer always corresponds to the same instantaneous velocity of the source. The whole velocity range of interest is scanned many (typically 10 – 50) times per second. As each instantaneous velocity occurs twice during one scan period, the spectrum is registered twice in the memory, where both halves of the memory appear as mirror images, symmetrically about the center of the channel series.

It is essential to have a standard reference line position which other signals can be referred to, particularly in Mössbauer spectroscopy, where different host metals for a particular Mössbauer isotope are often utilized, e.g., ^{57}Co/Cu, ^{57}Co/Pt, ^{57}Co/Rh in case of ^{57}Fe measurements with emitting γ-rays of slightly different energy. The most widely used reference material for velocity calibration in Mössbauer spectroscopy is a natural iron foil of which the six line positions of the magnetic hyperfine split spectrum are accurately known. One uses the distance between the two outermost lines of the iron spectrum as a calibration basis for the measurements of the hyperfine interactions with strengths corresponding to a Doppler velocity region of $0-10\,\mathrm{mm\,s^{-1}}$. This region is suitable for Mössbauer-effect measurements in, e.g., ^{57}Fe, ^{119}Sn, ^{61}Ni, ^{197}Au, and others.

The usual detectors for low-energy γ radiation are thin NaI(Tl) scintillation crystals coupled to a photomultiplier, and gas-filled proportional counters. Considerable improvements in energy resolution can be achieved with solid-state detectors such as the Li-drifted Ge counter.

It is only with low-energy γ emitters, such as the excited state of ^{57}Fe, that the Mössbauer effect can be observed at room temperature and even above. With most Mössbauer isotopes it is necessary to cool the absorber and sometimes also the source to liquid-nitrogen or liquid-helium temperatures to obtain a measurable effect. Apart from these practical requirements

it is often desirable to vary the temperature of the absorber in order, e.g., to obtain information about phase transitions or the electronic state and the molecular symmetry of the compound of interest. Many types of cryostats are nowadays commercially available. The temperature range varies from 1.2 to 300 K in He-bath cryostats and from 6 to 300 K in He-flow cryostats.

A He-bath cryostat equipped with a superconducting magnet can be utilized for Mössbauer investigations of materials in an applied magnetic field. Magnetic fields up to 10 T are common today.

Above room temperature electrical furnaces are used. To minimize the risk of oxidation of the material, a moderate vacuum of about $10^{-2} - 10^{-3}$ mbar is appropriate.

4. Preparation of Mössbauer Source and Absorber

The source for a Mössbauer experiment is a radioactive isotope of reasonable half-life and appropriate nuclear transition energy and excited state lifetime. By radioactive disintegration, the isotope populates a nuclear excited state which decays to the ground state by emitting low-energy γ radiation. Table 1 lists a selection of isotopes which are suitable for practical Mössbauer-effect measurements.

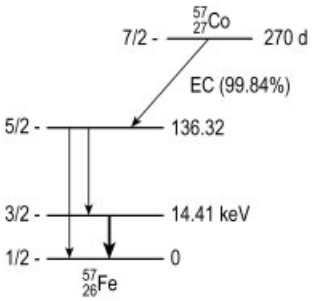

Figure 6. Decay scheme of ^{57}Co

As a typical example, Figure 6 shows the decay scheme of ^{57}Co which populates the 14.4 keV Mössbauer level of ^{57}Fe with a lifetime of $\tau = 140$ ns. The isotope ^{57}Co can be produced in a cyclotron by the nuclear reaction ^{56}Fe(d,n)^{57}Co. The decay of ^{57}Co occurs essentially by electron capture (99.8 %) from the K shell leaving a hole in this shell which is filled from higher shells with emission of a 6.4 keV X-ray. Sources of ^{57}Co are usually prepared by electrochemically depositing the carrier-free isotope on metallic supports and then diffusing it into the metal at high temperatures.

A Mössbauer source should meet the following requirements.

- The emission line should be as narrow and intense as possible, unsplit, and unbroadened
- The recoil-free fraction should be as high as possible
- The source material should be chemically inert during the lifetime of the source and resistant to autoradiolysis
- The host material should not give rise to interfering X-rays, and Compton scattering and photoelectric processes should be insignificant

A uniform absorber with randomly oriented crystallites can be prepared easily by sandwiching the finely ground material between the thin windows of a Perspex holder. In order to obtain undistorted line shapes it is desirable that the absorber is "thin". The natural line width Γ_{nat} of the Mössbauer line having Lorentzian shape [see Section 2.1, Eq. (3)] is determined by the half-life of the excited nuclear state and the Heisenberg uncertainty principle. The experimental line, however, is broadened (Γ_{exp}) for several reasons. The hyperfine fields may have a distribution which is typically observed in amorphous materials and small particles (nano particles). The narrow hyperfine field distribution caused by imperfections and strains in molecular crystals is typically not observed as a line broadening. Dynamic behavior of hyperfine fields lead to line broadenings if the eneries $\hbar\tau_{\mathrm{f}}$ corresponding to the fluctuation time τ_{f} of the fields at the nucleus are of the same order of magnitude as the difference of the hyperfine energies belonging to the field values. The sample thickness gives rise to the so-called thickness broadening which, however, can be taken into account by fitting routines on evaluating the spectra.

Heavy elements in the compound under investigation will scatter and absorb the Mössbauer γ-rays and/or emit photons of similar energy, which are detected along with the Mössbauer photons and lead to a decrease in the signal-to-noise ratio. Mössbauer lines of appropriate intensity can therefore be obtained by

Table 1. Some important isotopes in Mössbauer spectroscopy

Mössbauer isotope	Abundance of stable isotope, %	γ-Ray energy, keV	Half-life of Mössbauer transition, 10^{-9} s	Parent isotope	Half-life of parent isotope
^{57}Fe	2.14	14.4	97.8	^{57}Co	270 d
^{119}Sn	8.58	23.83	17.8	^{119m}Sn	245 d
^{121}Sb	57.25	37.15	3.5	^{121m}Sn	76 y
^{129}I	0 (β^-)	27.77	16.8	^{129}Te	33 d
^{151}Eu	47.82	21.54	9.5	^{151}Sm	90 y
^{197}Au	100	77.35	1.89	^{197}Pt	18 h

decreasing the absorber thickness and increasing the recording time. It is therefore preferable to estimate an optimum absorber thickness for which the measuring time is as short as possible in order to achieve the maximum information from the experiment [17], [18].

5. Hyperfine Interactions

Owing to the narrow linewidth of the Mössbauer nuclear transition (on the order of 10^{-8} eV), the resonance spectrum is extremely sensitive to energy variations of the γ radiation. Interactions between the nucleus and the surrounding electrons with energies comparable to the width of the transition lines manifest themselves in the Mössbauer spectrum. It is the influence of the electronic environment on the emission and absorption lines which determines the hyperfine structure of the spectrum. This interaction of the positively charged nucleus with the electric and magnetic fields caused by the orbital electrons in the region of the nucleus is called hyperfine interaction. This perturbation shifts or splits degenerate nuclear levels.

The Hamiltonian for the nucleus in a solid usually is expressed in a multipole expansion up to the second order

$$H = H_{E_0} + H_{M_1} + H_{E_2} + \dots \tag{12}$$

Higher order terms are negligibly small and cannot be detected by Mössbauer effect measurements. The first term, the so-called monopole interaction, E_0, gives rise to the *isomer shift* δ and shifts the energy levels of both the ground state and the excited state. It originates from the Coulombic interaction between the nucleus and electrons at the nuclear site. The magnetic dipole interaction M_1 removes the degeneracy of the nuclear levels if the nucleus has a magnetic moment. It is commonly known as the magnetic hyperfine interaction and describes the influence of a magnetic field on the nuclear spin (*magnetic splitting*). The third term in Equation (12) represents the electric quadrupole interaction, i.e., the interaction between the nuclear quadrupole moment and an electric field gradient at the nuclear site. The resulting Mössbauer parameter is called *quadrupole splitting*.

The different interactions and the kind of information one can obtain from the respective Mössbauer parameters are summarized in Table 2.

5.1. Isomer Shift

The isomer shift (other names occasionally used are chemical isomer shift or center shift) arises from the fact that the nucleus of an atom possesses a finite volume, and s-electrons have the ability of penetrating the nucleus and spending a fraction of time inside the nuclear region. *p*, *d*, and *f* electrons do not have this ability.

The electric monopole energy of a uniform spherical nucleus with radius and charge in a constant electron charge density $-e|\psi(0)|^2$ is given by

$$H_{E_0} = \frac{2}{5}\pi Z e^2 \left|\psi(0)\right|^2 R^2 \tag{13}$$

The nuclear radii of the ground and excited states, in general, differ by a small amount δR. This results in a gain or loss of Coulombic energy δE of the nucleus when it emits a γ quantum. This energy is transferred to the γ-ray and therefore produces a shift. The same argument holds for the absorption process, so that the total shift δ is expressed by

$$\delta = \frac{4}{5}\pi Z e^2 R^2 (\delta R/R) \left\{ \left|\psi_a(0)\right|^2 - \left|\psi_s(0)\right|^2 \right\} \tag{14}$$

Table 2. The interactions between the nucleus and surrounding electrons, the respective Mössbauer parameters, and the type of information

Type of interaction	Mössbauer parameter	Information
Electric monopole interaction between nucleus and electrons at the nuclear site	Isomer shift δ	a) Oxidation state (nominal valency) of the Mössbauer atom. b) Bonding properties in coordination compounds (covalency effects between central atom/ion and ligands. Delocalization of d-electrons due to back-bonding effects, shielding of *s* electrons by *p* and *d* electrons). c) Electronegativity of ligands.
Electric quadrupole interaction between electric quadrupole moment of the nucleus and electric field gradient at the nuclear site	Quadrupole splitting ΔE_{Q}	a) Molecular symmetry. b) Oxidation state (nominal valency). c) Spin state. d) Bonding properties.
Magnetic dipole interaction between magnetic dipole moment of the nucleus and a magnetic field at the nucleus	Magnetic splitting ΔE_{M}	Magnetic properties (e.g., ferro-, antiferro-, para-, diamagnetism; absolute value and direction of local magnetic fields).

where $\delta R = R_{\mathrm{e}} - R_{\mathrm{g}}$ is the change in nuclear radius on going from the excited (e) to the ground state (g). The subscripts a and s refer to absorber and source, respectively. Figure 7 shows schematically the origin of the isomer shift. From Equation (14) it is clear that the isomer shift is always a relative shift, between absorber and source.

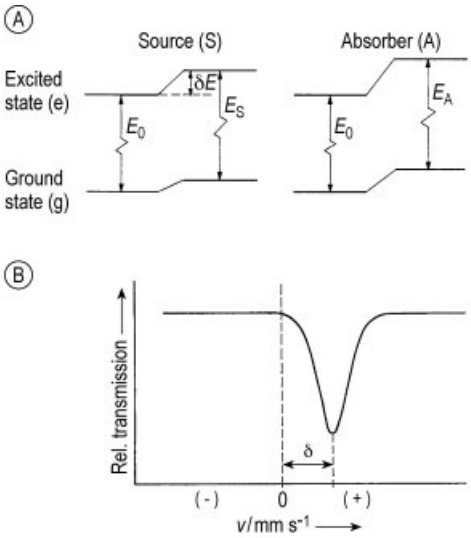

Figure 7. Origin of the isomer shift
A) Electric monopole interaction in source and absorber shifts the nuclear energy levels without affecting the degeneracy; B) Resultant Mössbauer spectrum (schematic)

Therefore, the measured isomer shift is always given with respect to a standard material; this can be the Mössbauer source used in the particular experiment or any conventional absorber material.

The isomer shift is expressed by a nuclear term $\delta R/R$ and an electronic term $\left\{ \left| \psi_{\mathrm{a}}(0) \right|^2 - \left| \psi_{\mathrm{s}}(0) \right|^2 \right\}$, which measures the electron density at the nucleus of an absorber relative to a given source. For a given Mössbauer atom the nuclear factor is constant; the isomer shift is thus exclusively dependent on the difference of the electron densities between absorber and source. For ^{57}Fe $\delta R/R$ is negative, i.e., the nuclear radius in the excited state is smaller than in the ground state; in the case of ^{119}Sn the reverse is true.

The isomer shift can reflect minute changes in the electron density at the nucleus. In a nonrelativistic approximation, the electron density at the nucleus is large only for electrons with zero angular momentum (*s* electrons) and can be approximated by $—\psi(0)—^2$. Hence, the isomer shift of the Mössbauer spectrum is a relative measure of the total *s*-electron density at the nuclear probe in a compound.

The total *s*-electron density of an atom in a chemical compound is the sum of all contributions from the inner (filled) and outer (valence electron) shells. The contribution from the outer electron shells is sensitive to changes in the chemical environment of the element (e.g., change of charge state, change of bonding properties by electron delocalization) and will exert a direct influence on the isomer shift. The valence *s*, *d*, and *f* electrons have only an indirect influence on isomer shift by their shielding effect

on the outer s-electron shells against the positive nuclear charge.

The measurement of the isomer shift gives information on

- The oxidation state (e.g., Fe^{II}, Fe^{III}, Fe^{IV}),
- The spin state (high spin, low spin),
- The bond properties (covalent, ionic),
- The electronegativity of the ligands

of a Mössbauer atom in solid material (coordination compounds, alloys, intermetallic phases, amorphous materials, etc.).

Figure 8 shows approximate ranges of the isomer shifts determined in iron compounds with different oxidation and spin states of the central metal ion.

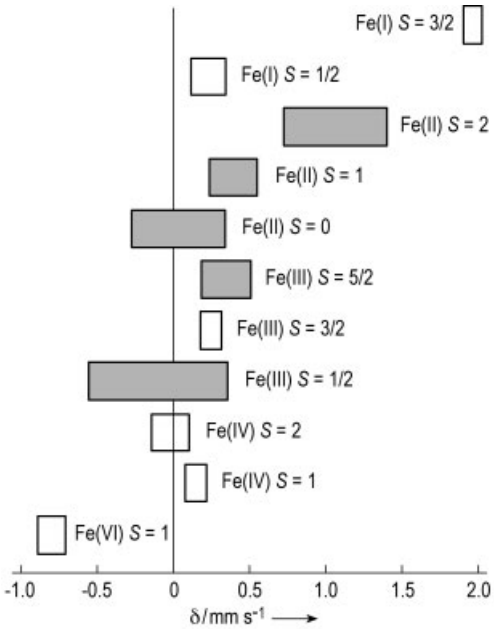

Figure 8. Approximate ranges of isomer shifts observed in iron compounds (relative to metallic iron). S is the electronic spin quantum number. Note that $(\delta R/R) < 0$ for ^{57}Fe.

It was shown [19] that the electronic density difference on going from the configuration $3d^6$ (Fe^{2+}) to $3d^5$ (Fe^{3+}) essentially originates from changes in the $3s$ shell. The filled $1s$ and $2s$ orbitals remain practically constant for both configurations. The removal of the $3d$ electron leads to an increase in the electron density at the nucleus due to the less effective shielding of the $3s$ electrons from the nuclear charge by the $3d$ electrons in Fe^{3+} as compared with Fe^{2+}. As $\delta R/R$ is negative for ^{57}Fe the isomer shift becomes more negative for Fe^{3+} than for Fe^{2+} (with respect to the same standard). The more or less widespread ranges of the δ values for each oxidation state is a direct consequence of the nature of the chemical bond. The electron distribution in the molecular orbitals is the result of the variable abilities of the ligands to donate electrons to the metal ion by σ bonding and to accept electrons from the metal ion by π bonding.

Correlation diagrams similar to that shown in Figure 8 are available for numerous Mössbauer atoms [20].

5.2. Magnetic Hyperfine Interaction

An atomic nucleus, either in the ground or excited state, with nuclear spin quantum number $I > 0$ possesses a magnetic dipole moment $\vec{\mu}$ and interacts with a magnetic field $\vec{B}$ at the nuclear site. The magnetic field $\vec{B}$ may arise from the electronic environment or from an external magnet. This interaction is called *magnetic dipole interaction* and is described by the Hamiltonian

$$H_{M_1} = -\vec{\mu}\vec{B} = -g_N \mu_N \vec{I}\vec{B} \tag{15}$$

where g_N is the nuclear Landé factor and $\mu_N = e\hbar/2Mc$ (M is the mass of the nucleus) is the nuclear magneton. The magnetic dipole interaction splits a nuclear state $—I\rangle$ into $2I + 1$ equally spaced substates $—I, m_I\rangle$, each of which is characterized by the nuclear magnetic spin quantum number $m_I = -I, -I+1, ..., I$ (nuclear Zeeman effect). The Zeeman energies of the substates are $—I, m_I\rangle$

$$E_M(m_I) = -\mu B M_I/I = -g_N \mu_n B m_I \tag{16}$$

Since the Mössbauer transition in ^{57}Fe is of the magnetic dipole type (M_1), there are only transitions between nuclear sublevels with $\Delta m_I = 0$, ± 1 and $\Delta I = \pm 1$. This selection rule yields only six transitions between the two ground-state sublevels ($I = 1/2$) and the four excited-state sublevels ($I = 3/2$). Figure 9 illustrates the splitting of the nuclear levels, the allowed transitions, and the resulting ^{57}Fe Mössbauer spectrum: a sextet due to the magnetic hyperfine interaction.

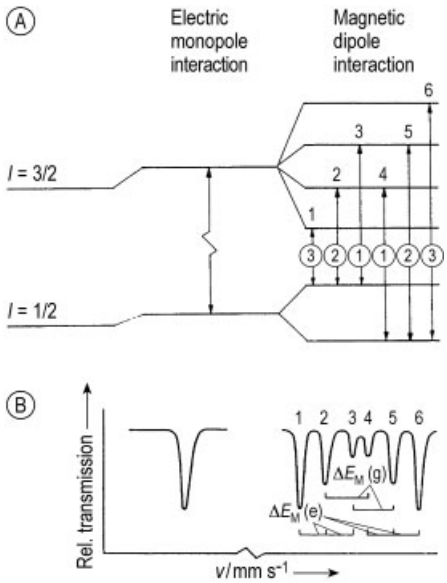

Figure 9. Origin of magnetic hyperfine splitting
A) Magnetic dipole interaction in ^{57}Fe in absorber; B) Resultant Mössbauer spectrum (schematic)

The magnetic hyperfine field $\vec{B}$ at the nucleus can originate from its own electron shell ($\vec{B}_{\mathrm{ion}}$), can be generated by paramagnetic ions in the nearby lattice ($\vec{B}_{\mathrm{lat}}$), or can be due to an applied magnetic field ($\vec{B}_{\mathrm{app}}$):

$$\vec{B} = \vec{B}_{\mathrm{ion}} + \vec{B}_{\mathrm{lat}} + \vec{B}_{\mathrm{app}} \tag{17}$$

The dipolar field of the moments of the surrounding paramagnetic ions $\vec{B}_{\mathrm{lat}}$ is a very small contribution.

$\vec{B}_{\mathrm{FC}}$ is the so-called Fermi contact field [21] and represents the most important contribution to the field $\vec{B}_{\mathrm{ion}}$. This field is the direct coupling between the nucleus and the unpaired *s*-electron density at the nucleus [22], [23],

$$\vec{B}_{\mathrm{FC}} = -\frac{16\pi}{3}\mu_B\langle\sum\left(\uparrow\psi_s^2(0) - \downarrow\psi_s^2(0)\right)\rangle \tag{18}$$

where $\uparrow\psi_{\mathrm{s}}^2(0)$ is the spin density at the nucleus of s-electrons with spin parallel to the total spins of the valence d-electrons spin and $\downarrow\psi_{\mathrm{s}}^2(0)$ that with spin antiparallel to the *d*-electron spins.

Since the exchange interaction between the spin-up polarized *d* shell and the spin-up *s* electron is attractive, while that between the spin-up *d* shell and the spin-down *s* electron is repulsive, the radial parts of the two s-electrons distort, one being closer to the nucleus and the other being more distant. As a consequence, the spin-down spin density is enhanced at the nucleus. The direction of the Fermi contact field is antiparallel to this spin, and consequently the Fermi contact field is negative by defintion, which means ant-parallel to the magnetic moments of the *d* electrons.

For high-spin Fe^{III} compounds with electron configuration $3d^5$ (^{6}S), the dipolar field due to its own electrons and the contribution of the orbital momentum to the total field are zero. The Fermi contact term, estimated to be -11 T per unpaired 3*d* electron, dominates over the contributions of the neighboring dipoles and the applied field. In high spin Fe^{II} compounds, however, the negative Fermi contact term is opposed by a positive orbital contribution of almost the same order of magnitude. The positive dipolar contribution is in general quite small. In this case it is impossible to predict the sign of the internal hyperfine field, B_{ion}. This problem can be solved by applying an external magnetic field that aligns the 3*d* moments. Sign and absolute value of the internal field can then be obtained from the vector sum of the applied and the internal field.

5.3. Quadrupole Splitting

The third term in the multipole expansion (Eq 12) describes the *electric quadrupole interaction*. This type of nuclear – electronic interaction is also electrostatic in nature and occurs if the following conditions are fulfilled:

1) The Mössbauer nucleus must possess a measurable electric quadrupole moment *eQ*, which arises from nonspherical nuclear charge distribution leading to spin quantum numbers $I > 1/2$. A measure of the deviation from spherical symmetry is given by the electric quadrupole moment *eQ* which can be calculated by the spatial integral

$$eQ = \int pr^2\left(3cos^2\theta - 1\right)dV \tag{19}$$

where *e* is the proton charge and *p* is the charge density at the spherical coordinates

r and θ. The sign of the nuclear quadrupole moment refers to the shape of the distorted nucleus: Q is negative for a flattened and positive for an elongated nucleus.

2) The electric field at the Mössbauer nucleus must be inhomogeneous, as measured by the electric-field gradient (EFG) $V_{ij} = \partial^2 V/\partial i \partial j$ ($i, j = x, y, z$). This arises from a noncubic charge distribution of the electrons and/or of the neighboring ions around the nucleus.

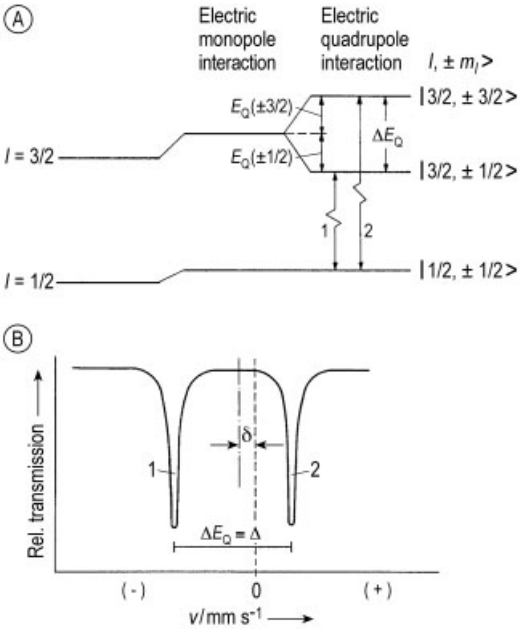

Figure 10. Origin of the quadrupole splitting in ^{57}Fe
A) electric quadrupole interaction in case of $eQV_{zz} > 0$; B) Resultant Mössbauer spectrum (schematic) with equal intensities as for a powder sample; the inevitable shift of the nuclear levels due to electric monopole interaction giving rise to the isomer shift is also shown

With respect to the principal axes, the EFG tensor is described by only two independent parameters, usually chosen as $V_{zz} = eq$ and $0 \leq \eta = (V_{xx} - V_{yy})/V_{zz} \leq 1$.

The spin Hamiltonian of the nucleus in the principal axes system of the EFG is given by:

$$H_{E_2} = \frac{eQV_{zz}}{4I\,(2I-1)} \times \left\{3\hat{I}_z^2 - I\,(I+1) + \frac{\eta}{2}\left(\hat{I}_+^2 + \hat{I}_-^2\right)\right\} \tag{20}$$

where $\hat{I}_{z,+,-}$ are the usual spin operators.

As an example, the effect of the electric quadrupole interaction in a Mössbauer nucleus with $I = 3/2$ in the excited state and $I = 1/2$ in the ground state, as is the case in ^{57}Fe and ^{119}Sn, is given in Figure 10. The quadrupole interaction gives rise to a doublet with the splitting

$$\Delta Eq = \frac{1}{2}e\left|QV_{zz}\right|\left(1 + \frac{\eta^2}{3}\right)^{\frac{1}{2}} \tag{21}$$

The quadrupole splitting in a Mössbauer spectrum is a highly sensitive probe for the chemical environment that determines the EFG at the nucleus of the Mössbauer atom. In general, there are two fundamental sources which can contribute to the total EFG.

1) Charges on ions surrounding the Mössbauer probe atom in noncubic symmetry give rise to the so-called *lattice contribution*, $(\mathrm{EFG})_{\mathrm{lat}}$ or $V_{zz}{}^{\mathrm{lat}}$ (for axial symmetry).
2) Noncubic distribution of the electrons in the partially filled valence orbitals of the Mössbauer atom generate the *valence contribution*, $(\mathrm{EFG})_{\mathrm{val}}$ or $V_{zz}{}^{\mathrm{val}}$ (for axial symmetry).

According to the different origins (valence/lattice contributions) it is expected that observed quadrupole splittings may reflect information about the electronic structure (oxidation state, spin state), bond properties, and molecular symmetry. As an example the spectra of three iron coordination compounds are shown in Figure 11.

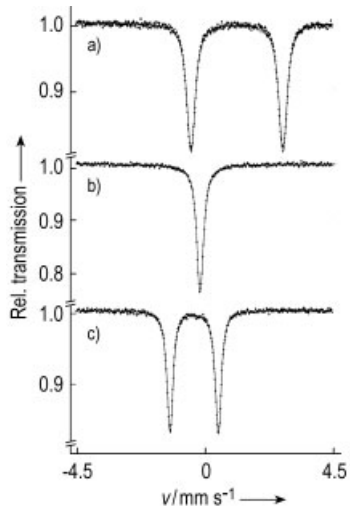

Figure 11. Mössbauer spectra of: a) $FeSO_4 \cdot 7\,H_2O$; b) $K_4[Fe(CN)_6] \cdot 3\,H_2O$; c) $Na_2[Fe(CN)_5(NO)] \cdot 2\,H_2O$

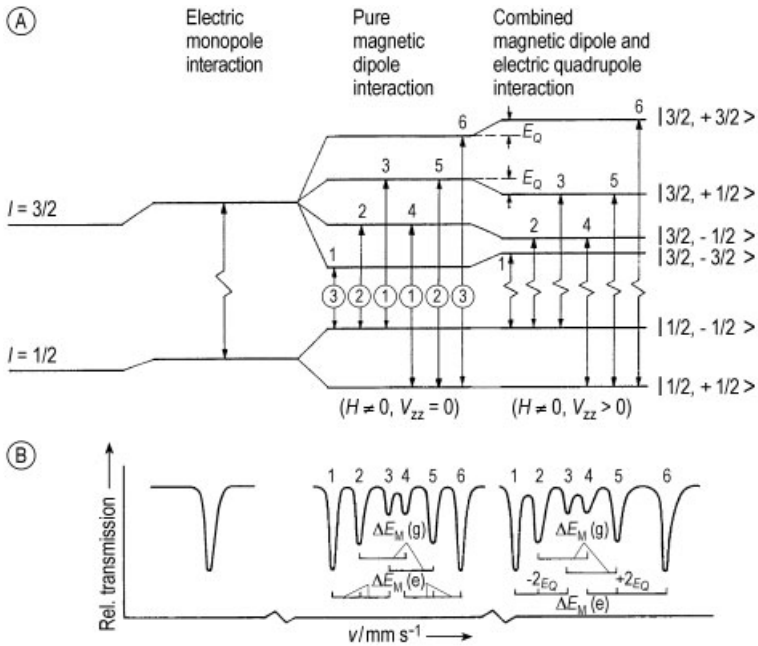

Figure 12. Combined quadrupole and magnetic hyperfine interaction
A) Splitting and shifts; B) Resultant Mössbauer spectrum (schematic)

1) The observed large quadrupole splitting of $\Delta E_Q = 3.4\,\mathrm{mm\,s^{-1}}$ for $FeSO_4 \cdot 7\,H_2O$ essentially arises from the valence contribution $V_{zz}{}^{\mathrm{val}}$. This compound is a typical high-spin compound in which the Fe(II) ion is octahedrally surrounded by six water molecules. This type of ion with a $^5T_{2g}$ ground state is subject to Jahn – Teller distortion, which lifts the degeneracy of the t_{2g} orbitals partly or completely and leads then to a noncubic distribution of the six electrons in the $3d$ shell of the $[Fe(H_2O)_6]^{2+}$ ions. The doubly occupied lowest d_{xy} orbital gives rise to the relatively large quadrupole splitting in the tetragonally compressed $[Fe(H_2O)_6]^{2+}$ octahedron. Further details can be found in reference [12].
2) $K_4[Fe(CN)_6] \cdot 3\,H_2O$ is a typical low-spin Fe^{II} compound. The single-line Mössbauer spectrum results from the absence of a quadrupole interaction due to the cubic charge distribution of the electron configuration $t_{2g}{}^6$ which consequently leads to zero valence contribution $V_{zz}{}^{\mathrm{val}}$ [12]. The lattice contribution $V_{zz}{}^{\mathrm{lat}}$ also vanishes since the $[Fe(CN)_6]^{4-}$ complex ion forms a regular octahedron.
3) The relatively large quadrupole splitting in the case of $Na_2[Fe(CN)_5(NO)] \cdot 2\,H_2O$ is solely due to the fact that the presence of different ligand molecules can no longer form a regular octahedron, as in the case of $[Fe(CN)_6]^{4-}$. The magnitude of the quadrupole splitting is mainly determined by the lattice contribution.

5.4. Combined Quadrupole and Magnetic Hyperfine Interactions

Quite frequently electric quadrupole interactions and magnetic dipole interactions may be present in addition to the electric monopole interactions which are always active. The total Hamiltonian H_{HFS} for the combined electric quadrupole and magnetic dipole interaction cannot be solved analytically in the general case of comparable interaction strengths and arbitrary direction of the magnetic hyperfine field in the PAS of the EFG. In case of a small EFG and large magnetic field, a situation often met, a perturbation treatment may be sufficient. For an axially symmetric EFG with the polar angle ϑ of the

direction of the magnetic field in the PAS of the EFG the eigenvalues of the nuclear sublevels of the $I = 3/2$ state are given by

$$E_{M,Q}(I, M_I) = -g_{\mathrm{N}}\mu_{\mathrm{N}}Bm_I + (-1)^{\left|m_I\right|+\frac{1}{2}}(eQV_{zz}/8)\left(3\cos^2\vartheta - 1\right) \quad (22)$$

We find in the case of ^{57}Fe, which is depicted in Figure 12, that the sublevels —3/2,± 3/2⟩ are shifted by an amount $E_Q(\pm m_I)$ to a higher energy and the sublevels —3/2,± 1/2⟩ are shifted by E_Q to a lower energy, if V_{zz} is positive. These energy shifts by E_Q are reversed if V_{zz} is negative. As a result, the sublevels of the excited state $I = 3/2$ are no longer equally spaced, unless $\cos\vartheta = \sqrt{1/3}$. Combined electric quadrupole and magnetic dipole interactions generally manifest themselves in the Mössbauer spectrum as an asymmetrically split sextet (for $I = 3/2 \longleftrightarrow I = 1/2$ transitions as in ^{57}Fe) as described in Figure 12. As the sublevel spacings and hence the asymmetry of the spectrum is directly correlated with the sign of V_{zz}, one can determine the sign of the EFG of a polycrystalline material from the magnetically split hyperfine spectrum. The effect of a magnetic dipole interaction (applied magnetic field) as a perturbation of the electric quadrupole splitting is described in detail in reference [24].

These three types of hyperfine interactions with the relevant Mössbauer parameters are most important in solid-state research. In addition, one often extracts further helpful information from the temperature and pressure dependence of the Mössbauer parameters, the shape and width of the resonance lines (relaxation phenomena), and the second-order Doppler shift (lattice dynamics).

5.5. Relative Intensities of Resonance Lines

The relative intensities of the lines in a Mössbauer spectrum provide additional information about the orientation of the hyperfine fields with respect to the crystal system in the case of single crystals or partially oriented (textured) samples. For pure magnetic or electric hyperfine fields the intensities are analytical expressions dependent on the direction of the fields with respect to the direction of γ radiation. For powder samples the intensity ratios of pure transitions do not depend on the strength of the fields.

For the case of a magnetically split spectrum of a 3/2 → 1/2 transition the intensity ratios are 3 : 2 : 1 : 1 : 2 : 3 for the hyperfine components of the Zeeman pattern. The corresponding quadrupole spectrum has equal intensities, as mentioned above. For more details the reader is referred to [25], [26].

6. Evaluation of Mössbauer Spectra

The evaluation of a Mössbauer spectrum implies the determination of the physical parameters, the isomer shift, Debye – Waller factor, and electric and magnetic hyperfine fields. When fluctuation times of the hyperfine fields of the order influencing the shape of the spectra are present the fluctuation times themselves may be a parameter of interest.

Fitting of Mössbauer spectra requires the existence of a model for the hyperfine interaction at the Mössbauer nuclei in the material under study. Computer programs are available for a variety of models. The interested reader is referred to the homepage of the Mössbauer community: http://www.kfki.hu/~mixhp.

7. Selected Applications

Since its discovery more than 40 years ago, Mössbauer spectroscopy has become an extremely powerful analytical tool for the investigation of various types of materials. In most cases, only two parameters are needed, i.e., the isomer shift and the quadrupole splitting, to identify a specific sample. In case of magnetically ordered materials, the magnetic dipole interaction is a further helpful parameter for characterization. In the following we shall discuss some applications selected from different research areas.

7.1. Magnetic Materials

As an example we will consider the magnetic ordering in magnetite, Fe_3O_4.

Magnetite is an inverse spinel, with iron atoms occupying interstitial sites in the close-packed arrangement of oxygen atoms. The tetrahedral sites (A) are occupied by ferric cations, and the octahedral sites (B) half by ferric and half by ferrous cations. In the unit cell there are twice as many B sites as A sites. The structural formula can therefore be written: $(Fe^{III})_{tet}[Fe^{II}Fe^{III}]_{oct}O_4$. There is little distortion in the cubic close packing of the oxygen atoms, so that the structure is close to ideal.

Magnetite has a Curie temperature of 840 K. In the magnetically ordered phase the iron atoms on A-sites are antiferromagnetically coupled to those on B sites, and this results in a complete cancelling of the moments of the Fe^{III} ions. Thus, a ferromagnetic moment, originating in the Fe^{II} ions on A sites, remains. On the whole, magnetite shows ferrimagnetic behavior.

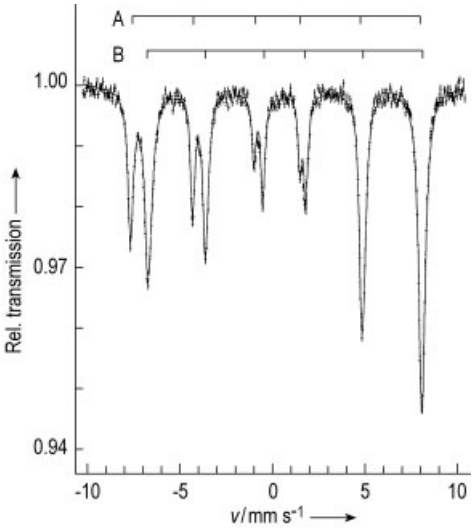

Figure 13. Room-temperature Mössbauer spectrum of magnetite; subspectra of lattice sites A and B are indicated by the stick diagram

The statistical distribution of the di- and trivalent ions of iron on equivalent lattice sites (B) explains many unusual properties of magnetite, in particular its extremely high electrical conductivity comparable to that of metals. Investigations have shown that the conductivity is caused by electron and not by ion transport. A fast electron-transfer process (electron hopping) between ferrous and ferric ions on the octahedral B sites takes place [27]. The Mössbauer spectrum of magnetite at room temperature (Fig. 13) confirms this interpretation.

The spectrum consists of two overlapping six-line spectra with internal magnetic fields of 49 and 46 T and a relative area ratio of approximately 1 : 2, which originate from the iron ions on sites A (Fe^{III}) and B (average field of $Fe^{II} + Fe^{III}$). The Mössbauer parameters δ, ΔE_Q, and H_{int} for the sextet stemming from the site-B ions are intermediate between those of oxidic Fe^{II} and Fe^{III} compounds; i.e., the iron ions on lattice site B are a mixed-valent system of Class III in the Robin – Day classification. At ca. 120 K, magnetite shows a so-called Verwey transition. The rate of the electronic hopping between the Fe^{II} and Fe^{III} ions on the lattice B sites slows down such that it becomes slower than the reciprocal of the Mössbauer time window. This causes a significant change in the magnetically split subspectra [28], [29].

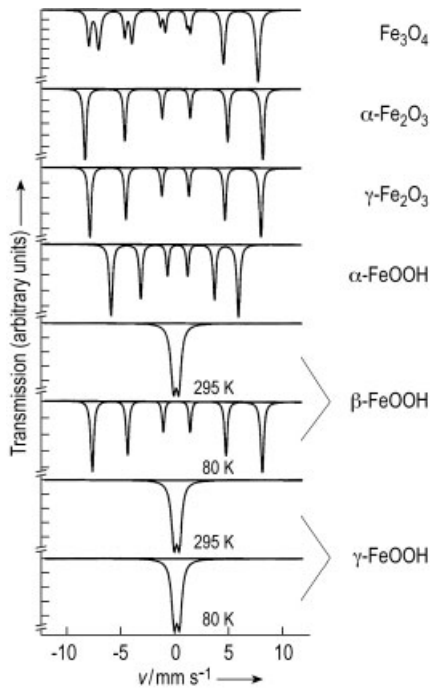

Figure 14. Mössbauer spectra of the most important iron-containing corrosion products

7.2. Corrosion Products

Mössbauer spectroscopy of ^{57}Fe has proven to be an extremely powerful tool for the analysis of iron-containing corrosion products on steel and other alloys. Depending on the corrosion conditions a variety of iron oxides and oxyhydroxides may be formed. Mössbauer spectroscopy enables one to differentiate between the various phases on the basis of their different magnetic behavior. Figure 14 illustrates this for six different corrosion products. Fe_3O_4 (magnetite) exhibits two superimposed six-line spectra, as described in Section 7.1.

α-Fe_2O_3 (hematite), γ-Fe_2O_3 (maghemite), and α-FeOOH (goethite) differ in the size of the internal magnetic field, which yields different spacings between corresponding resonance lines. A problem seems to arise with β- and γ-FeOOH, because they have nearly identical Mössbauer parameters at room temperature and cannot be distinguished by their Mössbauer spectra. However, lowering the temperature solves the problem: β-FeOOH appears to be magnetically ordered at 80 K, whereas γ-FeOOH still remains in the paramagnetic state.

Such Mössbauer studies may be carried out in a nondestructive manner and are even possible with highly dispersed particles which are no longer amenable to X-ray diffraction measurements.

7.3. Catalysis

In Figure 15 a series of spectra of the Fe/TiO_2 catalyst system [30] is shown. The top spectrum represents the freshly prepared catalyst; iron is present in the 3+ oxidation state, possibly in the form of finely dispersed oxide or oxyhydroxide on TiO_2. Under a reducing atmosphere (H_2 at 675 K) this iron phase is essentially converted into metallic iron (six line pattern). Some residues of Fe^{II} and Fe^{III} contribute to the spectrum as doublets. Under the conditions of the Fischer – Tropsch synthesis (FTS) at 575 K in CO and H_2 the metallic iron is transformed into the Hägg carbide χ-Fe_5C_2. The unreduced iron is now present as Fe^{II}, which is oxidized to Fe^{III} when the catalyst is exposed to air at ambient temperature. The carbide phase is left unchanged.

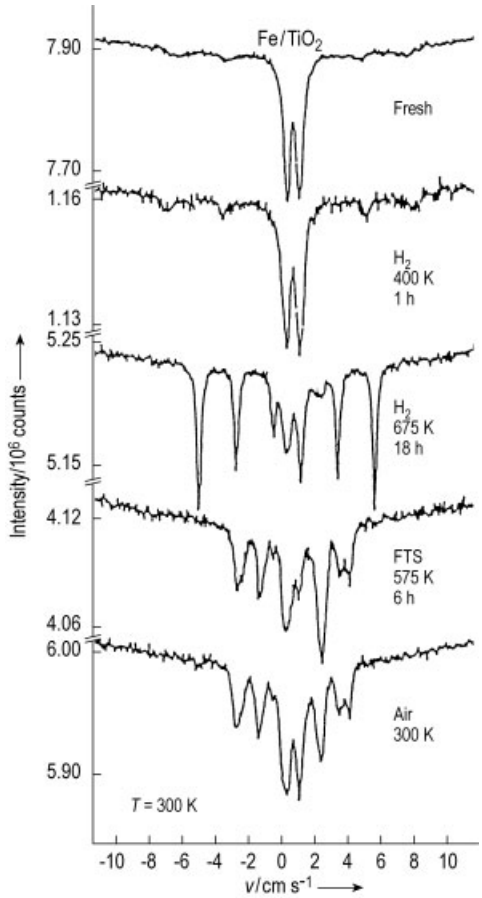

Figure 15. ^{57}Fe Mössbauer spectra at room temperature (from top to bottom) of the catalyst system Fe/TiO_2 after different treatments (see text)
Reproduced from *Applied Catalysis* [30]

7.4. Atmospheric Aerosol Samples

Iron is one of the most abundant elements in solid and aqueous atmospheric samples and is usually introduced into the atmosphere as soil dust, fly ash from power plants, exhaust from combustion engines, and from industrial operations. The role of iron in important atmospheric chemical reactions is subject to numerous investigations. Mössbauer spectroscopy proves to be a useful analytical method to determine the iron containing species in the atmosphere. In Figure 16 the result of the Mössbauer analysis [31] of an aerosol sample taken from an air-conditioning system of a building is shown.

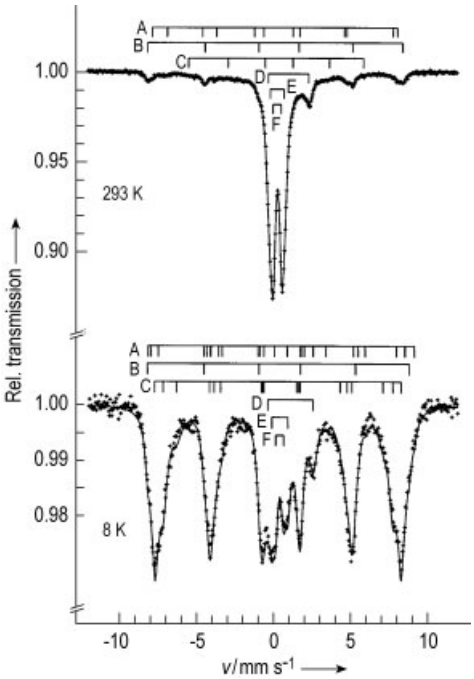

Figure 16. 293 K (top) and 8 K (bottom) Mössbauer spectra of the aerosol sample; the decomposition into the spectral components is indicated by the stick bars, where A stands for magnetite (Fe_3O_4), B for hematite (α-Fe_2O_3), C for goethite (α-FeOOH), D for Fe^{II} compounds, and E and F for Fe^{III} compounds

The typically very small crystals of the different components in the aerosol sample give rise to the so-called superparamagnetic (spm) behaviour (cf. Bronger et al. [32] and references therein) resulting in the pronounced resonance doublet for Fe^{3+} compounds in the room-temperature spectrum. The low-temperature spectrum (8 K) revealed that this doublet component mainly originates from superparamagnetic microcrystals of goethite and hematite.

7.5. U.S. One-Dollar Bill

Figure 17 shows the Mössbauer spectrum of a one-dollar bill. For the measurement the note was folded in such a manner that approximately 10 layers of the bill were analyzed. The result of the analysis: the iron-containing pigments of the banknote consist of about 20 wt % of magnetite and 80 wt % of goethite.

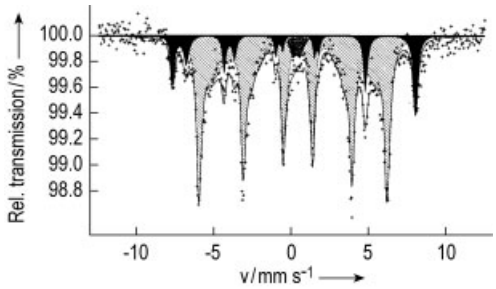

Figure 17. ^{57}Fe Mössbauer spectrum of a U.S. one-dollar bill. The different spectral components are marked as: dark gray = magnetite (Fe_3O_4), dashed lines = goethite (α-FeOOH), light gray = spm Fe^{III} component

8. References

1. R. Mössbauer, *Z. Phys.* **151** (1958) 124.
2. R. Mössbauer, *Naturwissenschaften* **45** (1959) 538.
3. R. Mössbauer, *Z. Naturforsch.* **14a** (1959) 211.
4. H. Frauenfelder: *The Mössbauer Effect,* Benjamin, New York 1962.
5. G. K. Wertheim: *Mössbauer Effect: Principles and Applications,* Academic Press, New York 1964.
6. H. Wegener: *Der Mössbauer-Effekt und seine Anwendung in Physik und Chemie,* Bibliographisches Institut, Mannheim 1965.
7. V. I. Goldanskii, R. Herber (eds.): *Chemical Applications of Mössbauer Spectroscopy,* Academic Press, New York 1968.
8. N. N. Greenwood, T. C. Gibb: *Mössbauer Spectroscopy,* Chapman and Hall, London 1971.
9. G. M. Bancroft: *Mössbauer Spectroscopy, An Introduction for Inorganic Chemists and Geochemists,* McGraw-Hill, New York 1973.
10. U. Gonser (ed.): "Mössbauer Spectroscopy," in *Topics in Applied Physics,* **vol. 5,** Springer, Berlin, Heidelberg, New York 1975.
11. T. C. Gibb: *Principles of Mössbauer Spectroscopy,* Chapman and Hall, London 1976.
12. P. Gütlich, R. Link, A. X. Trautwein: "Mössbauer Spectroscopy and Transition Metal Chemistry," in *Inorganic Chemistry Concepts,* **vol. 3,** Springer, Berlin, Heidelberg, New York 1978.
13. J. G. Stevens: Mössbauer Effect Reference and Data Journal, Mössbauer Data Center, University of North Carolina, Asheville, USA.

14. W. Kuhn, *Philos. Mag.* **8** (1929) 625.
15. D. Barb: *Grundlagen und Anwendungen der Mössbauerspektroskopie,* Akademie-Verlag, Berlin 1980.
16. E. Kankeleit: *Rev. Sci. Instrum.* **35** (1964) 194.
17. U. Shimony, *Nucl. Instrum. Methods* **37** (1965) 348.
18. S. Nagy, B. Levay, A. Vertes, *Acta. Chim. Acad. Sci. Hung.* **85** (1975) 273.
19. R. E. Watson, *Phys. Rev.* **118** (1960) 1036.
20. G. K. Shenoy, F. E. Wagner: *Mössbauer Isomer Shifts,* North Holland Publishing Company, Amsterdam, New York, Oxford 1978.
21. A. Abragam: *The Principles of Nuclear Magnetism,* Oxford University Press, Oxford 1961.
22. W. Marshall, *Phys. Rev.* **110** (1958) 1280.
23. W. Marshall, *Phys. Rev.* **105** (1957) 158.
24. R. L. Collins, *J. Chem. Phys.* **42** (1965) 1072.
25. T. Viegers: *Thesis,* University of Nijmegen, Netherlands 1976.
26. A. Gedikli, H. Winkler, E. Gerdau, *Z. Phys.* **267** (1974) 61.
27. E. J. W. Vervey, *Nature (London)* **144** (1939) 327.
28. R. S. Hargrove, W. Kundig, *Solid State Commun.* **8** (1970) 330.
29. M. Rubinstein, D. M. Forester, *Solid State Commun.* **9** (1971) 1675.
30. A. M. van der Kaan, R. C. H. Nonnekens, F. Stoop, J. W. Niemantsverdriet, *Appl. Catal.* **27** (1986) 285.
31. P. Hoffmann, A. N. Dedik, J. Ensling, S. Weinbruch, S. Weber, T. Sinner, P. Gütlich, H. M. Ortner, *J. Aerosol. Sci.* **27** (1996) 325.
32. A. Bronger, J. Ensling, P. Gütlich, H. Spiering, *Clays Clay Miner.* **31** (1983) 269.

Moth Control → Insect Control
Mouthwashes → Oral Hygiene Products
Multicomponent Polymer Systems → Polymer Blends
Muscle Relaxants → Skeletal Muscle Relaxants
Mustard Oils → Thiocyanates and Isothiocyanates, Organic

Meat and Meat Products

KARL-OTTO HONIKEL, Federal Center for Meat Research, Kulmbach, Germany

1. Definitions 1
2. History 2
3. Composition of Meat 2
4. Meat Production and Slaughtering 3
5. Fresh Meat Processing 5
6. Processing Techniques 6
6.1. Chilling of Meat 6
6.2. Freezing of Meat 7
6.3. Heating of Meat and Meat Products 8
6.4. Salting and Curing 8
6.5. Use of Additives 9
6.6. Smoking 9
6.7. Drying and Other Means of Preservation 10
7. Types of Meat Products 11
8. Manufacturing of Meat Products . 12
8.1. Raw Meat Products 12
8.1.1. Raw Hams 12
8.1.2. Raw Sausages 13
8.2. Cooked Products 13
8.2.1. Emulsion-Type Sausages 13
8.2.2. Cooked Sausages 14
8.3. Cooked Ham 14
9. Meat Species Identification 14
10. World Meat Production 15
11. Meat Grading 16
12. References 16

1. Definitions

The term *meat* means in a narrow sense all skeletal muscles of slaughtered warm-blooded, mostly domesticated animals. During the first days after slaughter the muscles are converted into meat (see Chap. 5) via chemical and physical processes.

Regulations in the Western world, however, summarize under the expression meat, all edible parts of a carcass [1], i.e., muscles with adjacent fat and fatty tissue — in some countries even organs such as liver, kidney, heart, brain and tongue, and also blood. Even if they are summarized under the term "meat", regulations demand different treatment for the various tissues (Chap. 6). Meat in a regulatory sense needs to be intact in structure and available in cuts larger than about 100 g [1]. Meat may only be chilled or frozen to be termed meat. Cutting to small cubes is already a processing step. Also no additions of any kind are permitted in meat. Some states outside the European Union permit the treatment of surfaces with antimicrobial substances (e.g., hypochlorite, ClO_2, H_2O_2, peracetic acid, organic acids) on fresh meat, especially poultry. Also in some states outside the EU the addition of a brine containing various salts (phosphates, $CaCl_2$) at a concentration of about 10 % in order to enhance tenderness and juiciness is permitted with declaration. The product is still called fresh meat.

Meat products and *meat preparations* designate all kinds of meat, either cut in small pieces or comminuted, and all cuts, small and large, covered with herbs or spices and/or treated with additions like salt, brine, or additives.

Skeletal muscles are termed cross-striated muscles, because of their appearance under the microscope (Fig. 1). The filaments (thick and thin) — their spatial arrangement are called myofibers — can slide into each other which causes a contraction (Fig. 2). The repeating unit is called a sarcomere. A sarcomere is about 1.8 – 2.0 µm long and may contract by 0.7 µm. The skeletal muscle contracts by nerve impulses setting Ca^{2+} ions free within the muscle cell. The energy is provided by hydrolysis of ATP to form inorganic phosphate and ADP. The chemical energy is converted into mechanical energy by inducing a series of conformational changes in the "head" of myosin in the thick filament, which binds to actin in the thin filament. After cessation of the nerve impulse, myofibers return into a relaxed state. Myosin is no longer bound to actin [2 – 4].

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.e16_e02.pub2

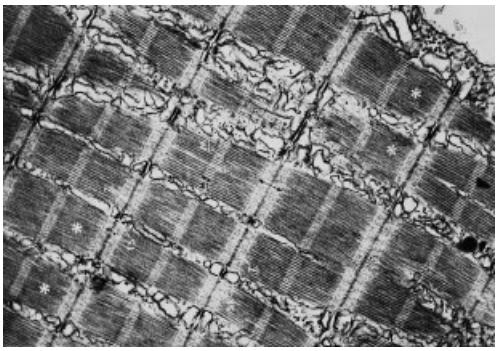

Figure 1. Micrograph of a cross-striated (skeletal) muscle

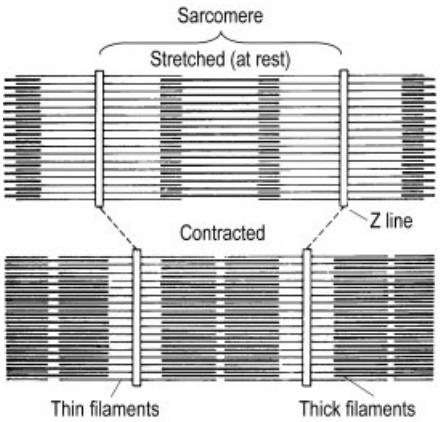

Figure 2. Scheme of a sarcomere of a cross-striated muscle

After death the energy is slowly exhausted, ATP disappears by hydrolysis to ADP and AMP and myosin binds permanently to actin. Rigor mortis occurs and the meat gets tough. It must age to become tender and tasty.

2. History

Since several thousand years wild animals have been domesticated. Besides dogs and cats cattle, goats, and sheep were the first animals which were kept by human beings. They produced milk and meat and cattle could be used furthermore for transporting goods or pulling ploughs and carriages. These variable uses demanded different handling. Female animals (cows) gave birth to calves and were used until they were old. Then they were slaughtered and meat, fat, and most organs were eaten. The skin was used for clothes or covers, the bones for tools or poles and the intestines were nearly ideal to keep meat in "containers" and preserve it. Only a few intact male animals (bulls) were kept for reproduction, most males were sterilized and used as steers for transport and pulling. Surplus male calves were slaughtered as unnecessary feed consumers. The obtained veal was tender, the meat of old cows, steers, and bulls was tasty but tougher. Cold storage of meat made it more tender (Chap. 3) but the storage required cooler climates or slaughtering in winter times only. Preservation techniques were developed (Chap. 8), namely drying, salting plus drying, or salting plus smoking. A limited preservation was also achieved by cooking.

The domestication of cows, sheep, and goats was followed by the keeping of chicken and pigs. Contrary to the former, which feed on grass, chicken and pigs were feeding on similar food as humans. Therefore pigs were not used as ubiquitous as cattle. Whereas cattle gave milk and meat for food, pigs produced mainly tasty fat besides meat. Fat of plant origin was available in sufficient quantities only in warmer regions. In moderate climates, such as Central, Western, and Northern Europe, pork fat was the main fat source besides dairy fat until the second half of the 19th century.

3. Composition of Meat

Lean meat is composed of about 75 % water, 22 % protein, 1 – 2 % fat, and about 1 % minerals and vitamins. In a carcass a wide variety of meat cuts with different compositions exist. In pork (Table 1) the fat content varies from 1.5 to 29 %. In beef (Table 2) the variations are nearly as wide. In poultry the maximum fat content is somewhat lower. In all main meat species eaten by humans the mean fat content of fresh meat cuts varies from 7.5 % fat for turkey, to 8.4 % for beef, 9.5 % for chicken, and 10.5 % for pork.

The protein of all meat species has a high biological value. More than 80 % of its amino acids can be used by the human body as the amino acid composition of meat is similar to human requirement [5]. The fat of slaughter animals consists mainly of monounsaturated fatty acids. The contents of unsaturated fat in fatty tissue is shown

Table 1. Composition of German pork cuts, g/100 g edible raw meat (N = 9 – 18)***

Cut	Protein	Fat	Fat		Energy content
	$\bar{\bar{x}}$	$\bar{\bar{x}}$	min	max	(mean) kJ (kcal)/100 g
Thick flank, knuckle	21.75	1.3	1	2	415 (99)
Topside	22.2	1.9	1.5	2.5	445 (106)
Tenderloin, fillet	22.0	2.0	1.5	2.5	445 (106)
Loin boneless	22.45	2.1	1	3	455 (108)
Chump	21.75	2.34	1.5	3	455 (108)
Loin bone-in, caudal (chops, cutlet)	21.6	7.5	3.5	14	645 (154)
Neck	19.7	9.6	7	11.5	660 (165)
Shoulder with rind	20.2	9.65	7	12	670 (168)
Loin bone-in, cranial (chops)	21.25	9.85	8	15	720 (174)
Forequarter hock, shank	20.35	10.8	9	11.5	752 (179)
Hindquarter hock, shank	18.95	12.2	9	14.5	780 (186)
Silverside with rind	19.2	15.35	11.5	17.5	903 (215)
Riblets	18.3	15.6	12	18	897 (214)
Shoulder blade	17.55	16.8	13.5	19.5	928 (221)
Belly, rindless	17.8	21.1	14	22	1097 (261)
Belly, rind-on	15.75	29.0	23	32	1361 (324)
Heart	16.9	2.1	2	2.5	363 (87)
Kidney	17.0	3.15	3	3.5	405 (96)
Liver	21.2	4.5	3.5	5	560 (133)
Tongue	16.4	15.7	13	17.5	869 (207)
Overall mean of meat cuts $\bar{\bar{x}}$ *	19.95	10.5			175/733
*S***	± 1.96	± 7.73			

* $\bar{\bar{x}}$ = Mean.
** s = Standard deviation.
*** N = number of samples investigated.

in Table 3. Lean meat cuts of all kinds contain around 5 % more unsaturated fatty acids than adipose tissue.

Meat also contains vitamins and trace elements (Table 4). Vitamins of the B group occur in high concentrations. Pork is very high in vitamin B1, 150 g of pork cover the recommended daily intake (RDI) of this vitamin. Lamb is high in vitamins B2 and B12.

Meat is low in sodium but high in potassium. Of the other mineral constituents of meat especially di- and trivalent ions such as iron are important as they are readily absorbed in the human intestines [8]. Due to the very balanced feeding of meat animals in the Western world also the level of trace elements like selenium or manganese is high in meat. Feed supplementation with vitamins, especially the lipophilic vitamins A and E, enhances their content in meat. Meat contains only minute amounts of vitamin C. Liver, however, has vitamin C concentrations of about 25 mg/100 g reaching the vitamin C concentrations of apples.

4. Meat Production and Slaughtering

Cattle, sheep, and goat are usually fed with grass, silage, and hay. Intensive feeding regimes of bulls and steers and many dairy cows require a high protein supplement with their feed. Most of the cattle are not kept in free ranging areas or raised in feedlots. They are kept inside, often due to climatic conditions.

Pigs and chicken are fed mainly with grain. They often receive a protein supplement in form of soybeans and other leguminoses. Most pigs are raised in-house due to climatic requirements. Free-range pigs are the exception. Poultry is also kept inside. Keeping of free-range poultry is rare but increasing.

In former times sheep, chicken, and pigs were slaughtered by cutting the throat. The stronger cattle has to be made unconscious (hammer on the forehead) or exhausted (by bulldogs) before killing.

In the beginning of the 20th century stunning (making unconscious) of animals became legally required in European countries. Captive bolt stunning was introduced, where a steel bolt penetrates the forehead by explosive devices. It

Table 2. Composition of German cuts of heifers, g/100 g edible raw meat (N = 6 – 15)***

Cut	Protein		Protein	Protein		Fat	Fat	Fat	Energy content
	$\bar{x}$[a]	±	s[b]	$\bar{x}$[a]	±	s[b]	min		(mean) kJ (kcal)/ 100 g
Eye of round	21.1	±	0.9	2.0	±	0.55	1.45		425 (102)
Rump	21.15	±	0.8	2.35	±	1.05	0.6		442 (106)
Thick flank, knuckle	20.8	±	0.7	2.85	±	0.85	1.9		455 (109)
Topside	21.8	±	0.9	2.6	±	0.95	1.35		463 (111)
Chuck tenderloin	19.9	±	1.1	2.45	±	0.95	1.6		425 (102)
Shank	21.35	±	0.55	2.35	±	1.1	1.35		446 (107)
Shin	20.9	±	1.0	2.5	±	0.9	1.5		442 (106)
Silverside	21.75	±	1.1	3.05	±	2.05	0.55		479 (114)
Tenderloin, fillet	21.2	±	0.7	4.0	±	1.55	2.05		506 (121)
Striploin, top muscle off	22.45	±	0.8	4.45	±	1.25	2.4		543 (130)
Shoulder clod	20.3	±	1.3	5.35	±	2.35	2.2		541 (129)
Sliced shank bone-in	20.3	±	1.8	5.35	±	1.8	2.55		541 (129)
Entrecôte	20.55	±	1.0	8.05	±	4.1	3.35		647 (155)
Chuck	19.0	±	0.5	8.85	±	1.3	7.3		651 (156)
Blade (shoulder)	18.25	±	0.85	8.85	±	2.2	5.3		639 (153)
Entrecôte to shoulder	19.25	±	1.05	9.25	±	1.4	6.9		682 (163)
Foreribs	19.4	±	1.45	10.6	±	3.7	4.1		724 (173)
Shoulder clod, tip	18.95	±	2.1	11.1	±	6.1	3.85		735 (176)
Silverside heel	18.15	±	1.05	12.25	±	3.3	6.75		765 (183)
Chuck to the loin	18.85	±	1.2	15.95	±	6.55	8.7		913 (218)
Brisket	18.35	±	0.9	14.45	±	3.1	10.7		851 (203)
Chuck	17.3	±	1.45	17.05	±	4.3	11.5		931 (223)
Plate	17.0	±	1.5	21.2	±	6.9	9.3		1078 (258)
Brisket tip	17.4	±	1.25	19.4	±	4.55	13.9		1022 (244)
Liver[c]	20.25	±	0.95	2.05	±	0.4	1.35		506 (121)
Overall mean, beef $\bar{x} \pm s$	19.7	±	1.6	8.4	±	5.9			644 (154)

[a] $\bar{x}$ = Mean.
[b] s = Standard deviation.
[c] Liver contains 5.3 % glycogen.
*** N = number of samples investigated.

Table 3. Unsaturated fatty acids in fatty tissue [6], % of total fatty acids

Species	Monounsaturated	Polyunsaturated	Total
Beef	43	5	48
Pork	46	11	57
Chicken	50	23	63
Turkey	27	29	56

Table 4. Vitamin and mineral content of heated meats [7]

		Vitamins, µg/100 g				Minerals, mg/100 g				
Species	Cut	B1	B2	B6	B12	Na	K	Fe	Mg	Zn
Beef	fillet	100	130	500	2.0	50	340	2.5	22	
Pork	fillet	900	230	500	5.0	60	420	1.0	18	1.9
Chicken	breast	70	90	–	0.4	70	260	1.1	20	1.0
Turkey	breast	47	81	460	05	50	330	1.0	20	1.8
Lamb	leg	150	370	130	2.7	67	290	1.6	22	2.9

is still the main way of stunning cattle in Europe. Later on stunning by passing an electric current through the brain was developed. Pigs and poultry are stunned today that way. A new technique is stunning by carbon dioxide. The animals are lowered into a CO_2 filled ditch where they turn unconscious. After stunning the animals are killed by debleeding usually by cutting blood vessels in the throat area (called sticking).

After sticking, when more than 95 % of the blood has left the carcass — about five minutes later — the carcasses are either cleaned by hot water or steam or are dehided, respectively defeathered. Then the belly is cut open and viscera are taken out.

After veterinary inspection the carcasses are weighed and graded with respect to meat and fat content (Chap. 11). The chilling process starts usually 30 to 45 min after stunning (Chap. 6). Chilling of cattle and pigs must be so effective that 7 °C are reached throughout the carcass at 24 to 36 h post mortem [9]. In practice pig carcasses are chilled to 7 °C within 16 to 18 h post mortem, cattle within 22 to 30 h. Poultry must be chilled to 4 °C as soon as possible usually within 2 – 6 h after death. Offals must be chilled to 3 °C. The temperatures must not rise above these limits until the carcasses are processed into products or prepared for consumption.

The carcass of cattle weighs about 55 – 65 % of the live weight of the animal [10]. In the case of pigs the carcass weight reaches about 70 %. From the total carcass weight the weight of the bones, tendons, and part of the non-eaten fat, in total about 15 – 20 %, must be deducted in order to obtain the edible portion. This means that about 50 – 60 % of the live weight of an animal is edible meat.

5. Fresh Meat Processing

During slaughter the bloodstream comes to an end. Oxygen and energy rich compounds are no longer transported to the cell and metabolic end products are not removed any longer. All tissues end their steady state metabolism sooner or later. Then they begin to disintegrate. Nerve cells stop their function about 20 min after death. Muscle cells switch from aerobic metabolism with CO_2 and H_2O as end products to an anaerobic breakdown of glycogen, ending with lactic acid [2, 11]. This means that only a small portion of the energy present in the form of carbohydrates (mainly glycogen) can be utilized post mortem. The anaerobic breakdown of glycogen to lactic acid produces three molecules of ATP (by reaction of ADP and phosphate) per glucose unit of glycogen. The aerobic breakdown to CO_2 and H_2O provides about twelve times as much ATP per glucose unit.

In a resting muscle about 5 – 7 mmol ATP per kilogram are available. 50 to 80 mmol glucose units of glycogen present per kilogram muscle can produce 150 to 240 mmol ATP per kilogram post mortem; 100 to 160 mmol lactic acid per kilogram meat are formed as the end product. With a pK_a of 3.9 of lactic acid a pH of about 2 could be reached theoretically in the muscle. But due to the buffering capacity of protein side chains containing free amino groups and other compounds, the pH in meat decreases only to 5.4 – 6.0 under normal conditions from about 7.0 in muscles of a living animal.

Fatty acids cannot be used as a source of chemical energy in post mortem muscle as oxygen is required to form acetyl-CoA. Oxygen is not available post mortem.

Another immediate energy source is creatine phosphate which buffers the ATP concentration in muscle cells. In a resting muscle ATP + creatine form creatine phosphate + ADP. When ATP is needed the reaction proceeds backwards.

Furthermore ADP can be disproportionated enzymatically (catalyzed by adenylate kinase or myokinase) according to

$$2\ \mathrm{ADP} \rightleftarrows \mathrm{ATP} + \mathrm{AMP}$$

ATP returns into the energy cycle, AMP is further metabolized

$$\mathrm{AMP} \rightarrow \mathrm{IMP}\ (\text{inosine monophosphate}) + \mathrm{NH_3}$$

During the early post mortem period energy rich compounds are broken down. Creatine phosphate forms creatine, glycogen is metabolized to lactic acid, ATP is hydrolyzed to ADP and AMP, which is deaminated to IMP. The latter and lactic acid are sensorically very favorable compounds. Both give flavor to the meat. The decrease in pH is of further advantage. It stops or retards the growth of many spoiling or hazardous microorganisms and thus prolongs shelf life.

From a structural point of view, the lack of energy-rich compounds initiates a selfdeteriorating process of the cells. With falling pH the well ordered compartmentation of the cells gets lost. Proteolytic and lipolytic enzymes begin to break down polymeric structures [12]. Especially the well-ordered protein chains in the skeletal muscles (see Chap. 1) are broken down to peptides and finally to amino acids. The loss of the ordered state leads to a tenderization of meat. The activity of proteases varies between muscles but primarily between species.

Tenderness/toughness of meat is determined by the myofibrillar structures of the filaments and the connective tissue [13]. The myofibers are the primary points of action of proteases post mortem. The connective tissue situated outside the muscle cells is attacked only after the cell walls become permeable for proteins. This happens 6 – 10 days post mortem, depending on the species.

Contrary to myofibers the structure of connective tissue depends on age and gender of the animals. Older and male animals develop more cross-links between the protein threads. This makes the connective tissue stronger (tougher).

Meat with strong connective tissue cross-links must age longer for the proteases to become effective. Meat from broilers and pigs — both have not yet reached adult age on slaughter — needs a shorter time of aging then that of cattle. Table 5 shows the minimum time of aging at 3 °C.

Table 5. Time of aging required for tender meat stored at + 3 °C

Species or gender	Time of aging, d
Chicken	1 – 2
Pork	3 – 4
Veal	7 – 8
Heifer meat	13 – 15
Young bull meat	17 – 21

The times in Table 5 are valid for intact meat, intended to be fried or grilled. Meat which is cooked or comminuted can be used much earlier. For Hamburgers (ground beef) meat may be used at two days post mortem.

Besides aging other processing steps influence tenderness. All processes that mechanically destroy structures, such as mincing or pressing, tenderize meat. Also the heating procedure is important. When the temperature rises beyond 40 °C, proteins denature. Fibrous proteins like the ones in myofibers and connective tissue shrink on denaturation [14], which means that the structure gets denser. Myofibrillar proteins start to shrink at around 45 to 50 °C: the meat becomes tougher. At a similar temperature range the heat labile cross-links of connective tissue are hydrolyzed which makes the meat tender. Around 65 °C the connective tissue shrinks. This is the temperature above which the meat gets tough again.

Proper heating is a complicated matter. In veal, pork, or chicken meat the heat-induced changes of toughness are small and a sufficient heating to 70 to 80 °C in order to kill microorganisms is appropriate and acceptable for tenderness. These meats with heat-labile cross-links are tender from around 55 °C (bloody, blue, sanguine) onwards. Heifer meat is most tender around 58 °C, young bull beef at 60 – 62 °C (medium done). Especially young bull beef increases in toughness again on heating above 62 °C up to 80 °C.

6. Processing Techniques

6.1. Chilling of Meat

As mentioned in Chapter 4 meat must be chilled to temperatures low enough to prevent the growth of spoiling or pathogenic microorganisms. Most of the bacteria stop growing below 8 – 10 °C [15]. Thus meat must be chilled to 7 °C as soon as practicable and possible. It is a legal requirement that beef, pork, and sheep meat must be chilled to 7 °C before transport.

Most hygienic rules require that the chilling process of slaughtered carcasses should commence not later than 45 min post mortem. At this time the eviscerated carcasses have been weighted and graded (Chap. 11). The temperature at the surface is below 30 °C but already 1 cm inside the carcass still body temperatures of 38 to 39 °C prevail. If stress has occurred during slaughter, e.g., in the case of pigs, inside temperature may even rise above 40 °C at 45 min post mortem.

Carcasses of cattle, pigs, and sheep are cut along the backbone into two halves. The halves hang on a rail fixed on their hindlegs head down. Cattle and sheep are dehided and the heads are

cut off. Pigs are hanging hide-on with their heads down. The pig carcasses have been scalded and dehaired. Poultry is eviscerated and defeathered, hanging on the rail as whole carcasses with spread legs.

Larger carcasses often pass at first through a freezing tunnel (blast chiller) for 45 to 90 min at temperatures of − 10 to − 20 °C where they are blown with cold air (air speeds up to 3 m/s) in order to remove the heat from the surface as fast as possible. This method has the advantage of reducing the evaporation loss of a carcass (up to 1 % within the chilling down period) and of reducing the time of chilling to 7 °C. Chill room throughput can be enhanced by using a blast chiller in the beginning.

The disadvantages of using a freezing tunnel early post mortem is that a muscle contraction may occur (cold shortening) [13]. Cold shortening takes place in the prerigor state when ATP is present in sufficient concentrations (Fig. 3). That is the case in beef and sheep until 10 h post mortem, in pigs until 4 h, in chicken and turkey until 45 min post mortem. Cold shortened meat gets tough and loses more fluid (drip loss, Fig. 4).

Figure 3. Shortening of beef muscles. Both muscles were excised at 45 min post mortem and cut to the same length, the top sample was chilled for 24 h at 0 °C, the bottom sample to 20 °C for 24 h.

Blast chilling may cause shortening in thinner parts or surface parts of muscles. If the blast chiller is followed by a chiller which keeps the temperatures in the already cold parts below + 10 °C, cold shortening prevails. In chill rooms with moderate chilling the temperatures inside the muscles (up to 40 °C) may warm up the surface above 10 °C and higher and release the muscle contraction. In consequence the temperature fall in a carcass must observe the rule: Fast but not too fast. Table 6 lists the best chilling conditions.

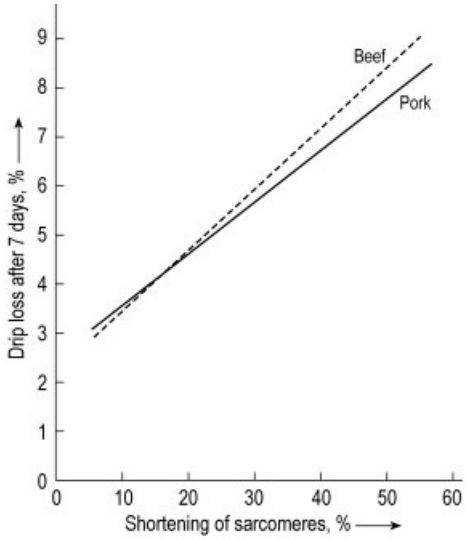

Figure 4. Drip loss of beef and pork muscles in dependence of shortening of sarcomeres

Table 6. Chilling regimes for pig carcasses

Normal chilling	In chiller: air at 0 – 4 °C; slow to medium air movement
Rapid chilling	freezer tunnel at − 10 °C, from 45 min to about 2.5 h post mortem at high air speed (up to 3 m/s and 95 % relative humidity) then 0 – 4 °C in chiller
Shock chilling	freezer tunnel at − 20 °C, from 45 min to 100 min post mortem (air conditions as with rapid chilled) then 0 – 4 °C in chiller

Another disadvantage is that a part of the meat freezes in the blast freezer. Freezing and thawing reduces the quality of meat and enhances drip and evaporation loss.

Instead of using a freezing tunnel, the chilling speed can also be enhanced by chilling with water mist or droplets. Due to the temperature of the water, temperatures of 10 °C and higher are maintained in the prerigor meat. The evaporation loss is minimized. The water vapor or droplets should, however, not be recycled in order to prevent cross contamination.

6.2. Freezing of Meat

Meat can be frozen after it has reached its final pH value. Prerigor meat either contracts before freezing or after thawing and becomes tough.

On freezing ice crystals form. Freezing occurs at first in the extracellular space due to lower concentrations of ions and functional groups of proteins (amino, hydroxy, or carboxyl groups). During slow freezing water moves from the intracellular to the extracellular space. Large ice crystals are formed penetrating the cell wall and enhancing the water migration. Thawing leads then to an enhanced drip loss. Fast freezing results in less migration of water and small ice crystals.

Freezing and thawing must be carried out at similar velocities to reduce drip loss. High velocities are better than lower ones, i.e., freezing of small pieces is advantageous. But in all cases drip loss is higher than in chilled meat. Perforation of cellular structures by ice crystals also eases the penetration of oxygen leading to faster development of off-flavors.

6.3. Heating of Meat and Meat Products

As described in Chapter 5, heating of meat causes structural changes leading to tough or tender meat. The prerequisite to obtain a tender heated piece of meat is the proper aging time (Table 5). Beyond this the heating regime can enhance or lower the tenderness [14]. The heating regime becomes important for meat with a strong connective tissue, i.e., beef, especially from bulls, and meat from older animals like cull cow meat, sow meat, mutton, or meat from hens. In beef the influence is most pronounced (see Chap. 5). Sanguine (bloody) or medium done beef is clearly more tender than well done beef. Cooking of beef above 80 °C and boiling in fluid for a longer period of time tenderizes the meat as the connective gelatinizes.

Heating of meat products fulfills the primary purpose of prolonging shelf life by killing microorganisms. Filling in casings or cans protects against recontamination.

In emulsion-type sausages and cooked hams furthermore a heat stable gel is formed during heating which permits the slicing of the products. Heating to ≥ 65 °C destroys microorganisms and induces gel fromation. Tenderness is not a major problem in comminuted and salted products. Here, the "bite" varies due to salt and additive concentration, the content of lean meat and fat and temperature of heating. The bite is determined by the processor by adjusting the processing conditions.

6.4. Salting and Curing

Salting is defined as the addition of sodium chloride, sometimes contaminated by small amounts of other salts. The first reason of salt addition is the reduction of the water activity (a_w) in order to inhibit or retard the growth of microorganisms. At $a_w < 0.95$ bacteria cannot grow [15], in products with $a_w < 0.7$ also molds stop growing. The second reason is flavor. Curing is the addition of salt plus nitrate or nitrite, respectively. Nitrite inhibits the growth of further microorganisms, e.g., *Clostridium botulinum* or enhances the intensity of the inhibiting action of salt [15]. Furthermore, in the presence of H^+ ions (pH $<$ 6.0) nitrite is converted to nitrous acid

$$NO_2^- + H^+ \rightarrow HNO_2$$

which forms a unstable anhydride

$$2\,HNO_2 \rightleftarrows N_2O_3 + H_2O$$

The anhydride disproportionates into NO and NO_2. NO binds to myoglobin, forming a heat-stable and red NO – myoglobin complex. Usually about 100 – 150 mg/kg nitrite are added with the salt. Due to food law, in several countries nitrite can be added only in a premix with salt, containing about 0.5 % nitrite, the so-called nitrite curing salt. If 2 % of curing salt is added to a product, 100 ppm nitrite are mixed in. If the product is heated, about half of the nitrite is metabolized during heating. The main part is oxidized to nitrate, i.e., nitrite acts additionally as an antioxidant due to scavenging of oxygen. During storage a further part of nitrite disappears as nitrate. Usually a heated meat product at the moment of consumption contains < 30 mg/kg nitrite [16]. In cases of lower pH values and temperatures above 150 °C in a product, nitrous acid forms a NO^+ cation

$$HNO_2 + H^+ \rightleftarrows NO^+ + H_2O$$

which reacts with secondary amines to form stable nitrosamines ($R_2N{-}NO$), that are carcinogenic. Under normal conditions of processing meat into heated meat products the pH is not low enough ($>$ pH 5.6), the temperature of the

heated products is below 100 °C and fresh meat has low concentrations of free amines. Thus, nitrosamines cannot form [17]. In raw meat products, stored for a long period, the content of amines increases and the pH is below 5.6. However, due to low temperatures of storage (ambient and lower) nitrosamines are formed at the limit of detection (ca. 1 µg/kg). Only in fried products (e.g., bacon prepared for breakfast) higher concentrations of nitrosamines occur.

Nitrate itself is ineffective beyond its ionic strength as an antioxidant and as an antimicrobial agent. It must be reduced to nitrite. Reduction takes place in fermented meat products. From the available 300 to 600 mg nitrate per kg, nitrite is slowly formed by special microbial action.

Salting and curing of meat pieces can be done by dry salt addition, immersing into salt brines, or injection of salt brines. During *dry salt curing* the meat pieces are fully covered with salt for several weeks in a chilled room. The salt penetrates slowly into the meat. After that time the surface is cleaned from salt and the product is dried, sometimes cold smoked. After a loss of about 30 % of weight, the product is ready for consumption. This drying period may last up to 12 – 24 months in traditional products. Usually it takes about 1 – 2 months of drying. The *immersion in brine* principally follows the same procedure. By *injection of brine* (20 to 30 %) a concentrated brine solution is pressed in the meat by a multi-needle injection. The following steps are equal to the other procedures. In minced products (all sausages) the salt is added during comminution. It causes swelling and dissolving of myofibrillar proteins which form heat-stable gels in emulsion type products (see Section 8.2.1).

6.5. Use of Additives

According to EU regulations many additives are permitted in meat products [18]. In reality, however, only a limited number of substances are used. The additives are applied primarily for technological reasons. In heated products they shall prevent the cookout of fat and/or water (jelly). Additives are used to improve the cohesion of particles or permit thin slicing of products. Some like nitrite and nitrate retard spoilage and additionally give the products a stable red color which deteriorates only with spoilage. Others lower the pH value of the batter, some are used as flavor enhancers, such as glutamate. Some countries also permit colorants [19].

In some cases the European regulations limit the amounts of additives which may be used. But other additives, especially those with GRAS status, are permitted in "quantum satis" amounts. A few substances are allowed just for the treatment of the surfaces in order to prevent the growth of molds or microorganisms at the outside of casings. One of those is smoke which does not only impart flavor but also contains antibacterial and antimycotic compounds.

Additives may act during processing. For instance diphosphates create a quasi prerigor state of the myofibers (i.e., they separate actin from myosin heads, see Chap. 1). This status enhances gel formation. Due to phosphatase activities in the meat batter diphosphate is hydrolyzed to monophosphate. The latter does not separate actin and myosin. All phosphates, however, sequester di- and trivalent cations and thus prevent the catalytic oxidation of SH side chains of cystein to form S–S bonds of cystin. Early oxidation of SH groups lowers the gel stability of heated batters.

Some high molecular additives like soy, other plant, and milk proteins also strengthen the gel stability as they form a heat-stable network with meat proteins.

6.6. Smoking

Smoke is formed when wood is burned in the presence of insufficient oxygen. Temperatures between 400 and 550 °C should prevail in the wood. Under these conditions pyrolysis of cellulose and lignin occurs [20]. Besides formation of CO_2, CO, water and some volatile compounds (e.g., alkanes) about half of the weight of wood is converted to the actual smoke. The smoke is released into the smoking chamber where sufficient oxygen is available. Secondary pyrolysis products are formed which give the smoke its aromatic flavor. These products include organic acids (e.g., acetic acid), aldehydes (formaldehyde and others), alkohols, phenols, polycyclic aromatic hydrocarbons (PAH)

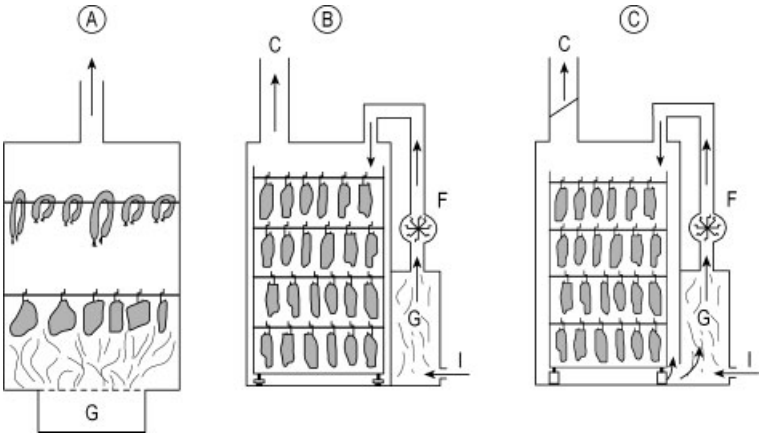

Figure 5. Smoking chamber
A) Traditional system; B) Modern open system; C) Modern closed system
G = Smoke generator by burning wood; I = Air inlet; F = fan; C = chimney

and other cyclic compounds. Since some of the latter may be carcinogenic, the PAH formation should be prevented or at least minimized. Temperatures around 500 to 600 °C create a minimum of PAH but a high concentration of flavor compounds. In 2005 the EU set MRL for PAH in food [21].

The smoke is introduced into a chamber where meat products hang freely on rails in order to permit the smoke to condense everywhere on the surface. Several types of smoke are applied: *cold smoke* (15 – 25 °C) usually applied on raw hams or raw sausages; *warm smoke* (25 – 50 °C) used for specialty products, e.g., for some emulsion-type product; and *hot smoking* (50 – 90 °C), which is often combined with a cooking process for heated meat products. The time of smoke application varies between 10 and 60 min.

Contrary to former times where burning and smoking was done in the same chamber (Fig. 5 A), nowadays the smoke is produced in a separate burner, entering the smoking chamber through a pipe system (Fig. 5 B and C). On its way the smoke will cool down and substances condense. The smoke may pass through the smoking chamber for a short period of time and leaves through a chimney. In a residential environment an afterburner must burn the smoke to CO_2 and water in order to prevent smell and odor. In modern systems smoke circulates within the chamber for quite some time before it is released into the afterburner. This method saves energy.

Due to carcinogenity of some smoke ingredients some countries limit the concentrations of the most hazardous PAH, benzo[a]pyrene. The concentration is limited by applying fresh smoke only up to 1 µg per kilogram of smoked product (1 ppb) [21]. Modern smoke technologies can easily follow this requirement.

Liquid smoke is a product which has been produced for 20 years. Freshly produced smoke is absorbed in edible oil or water under defined conditions. The liquid smoke or smoke flavoring is analyzed to ensure low or no contents of PAH and applied as an aerosol on the surface of meat products or as an ingredient. Liquid smoke is easy to handle and creates no odor problems in the environment. Furthermore due to the possible analysis prior to application it is low in or free of carcinogenic substances.

6.7. Drying and Other Means of Preservation

The first technologies of meat processing employed were heating and drying. For drying cool and dry climates were necessary. The process could be sped up by drying over a fire. In this way smoking was invented. Drying preserves by lowering the water activity.

Today, drying of meat is done carefully as *freeze drying*. Thin slices (2 – 4 mm thick) of

meat are frozen to − 20 °C and put in a vacuum chamber where they are kept frozen until most of the water has evaporated. Very brittle pieces of meat with 4 – 6 % water are obtained. Due to its brittleness the dried meat is often used in form of dried spherical particles (about 2 cm in diameter) or a coarse powder is made of the end product. The product must be produced in thin sheets or small pieces as the still existing cellular structures, i.e., phospholipid bilayers retard rehydration. Even a short cooking does not permit a rehydration of larger particles.

Another possible means of meat preservation is *irradiation* [22]. It is known from extended experiments that irradiation with 10 kGy is sufficient for preservation. But there is no country which permits the general use of irradiation for meat even if the WHO sees no problems. Only in food for special purposes (for sick or immunosuppressed people) the technology is used also for meat. In some countries the discussion continues if irradiation could be used to inactivate trichinellae.

A technology still rarely but increasingly used is the application of *high pressure*. 3000 to 10 000 bar are applied for about 2 min [23]. With pressures of 3000 – 5000 bar pathogens are inactivated. Lactobacilli survive partially these pressures. The most probable reason for inactivation is that cellular structures are disrupted by pressure application. In practice the vacuum packaged meat is put in a chamber and is surrounded by fluid. The pressure rises in about 2 – 3 min and is released in the same time interval. Due to structural changes meat turns gray and looks like cooked meat. Nitrite-cured meat products, however, keep their red color.

7. Types of Meat Products

Fresh meat is very perishable. Microorganisms can grow easily feeding on the constituents of meat. The reasons are manifold. One is the high water activity ($a_w > 0.99$), which means in simple words that more than 99 % of water are available as solvent and transport medium. Reducing the water activity retards or stops microbial growth. The most obvious way to do this is by drying (Section 6.7). But also higher concentrations of, e.g., salt or sugar lower the water activity (Section 6.4). Another way of preserving is the smoking of meat products as smoke contains antibacterial compounds (Section 6.6). Lowering of pH also retards spoiling microorganisms. Finally heating and subsequently protected storage in casings or cans enhances shelf life (Section 6.3).

Meat products have been invented primarily to prolong shelf life of meat. All aforementioned methods and their combinations are used. The most important meat products include:

1) *Dried Meat.* Thin slices of meat are dried in arid areas like prairies or arctic coasts, this is often preceded by salting. The preparation nowadays takes a few days only. Examples of dried meats are jerky beef (US Indians) or Biltong (South African natives). Dried meat is stored at ambient temperatures.
2) *Raw Meat Products.* Meat pieces are salted in a cool environment (< 15 °C) by covering with salt or immersing in salt brines. This step is followed by drying in dry air and partially by smoking. Examples without smoking include: Parma ham, Jamon Iberico, and Bündner meat. Time of processing is six months to two years, storage at ambient temperature. Examples of smoked products include: Schwarzwälder Schinken (Black Forest ham), Katenrauchschinken, and many other raw hams. Time of processing is one week to several month; the products are stored at ambient temperature.
3) *Raw Sausages*. Slightly frozen meat is comminuted with 2.5 to 3 % salt, spices and fat are added. Often starter cultures are mixed in. The cold batter (below + 5 °C) is filled in casings and incubated at 20 – 25 °C in order to permit the fast growth of the microflora of the starter cultures which lower the pH value. After 1 – 2 days the product is stored at 10 – 15 °C in a dry environment with air-movement. Some of the products are smoked. The product loses water (lowering the water activity) in the dry environment and is stabilized by higher salt concentration (ca. 3 % and above) and a low pH value (pH 5 – 5.5). The product is ready for consumption between one week and six months of processing, depending on the product. Dried products can be stored at ambient temperatures.
4) *Cooked Hams.* Meat pieces are salted as described in item (2). After salting the product is

heated to ca. 65 – 70 °C. The processing time is several days; cooked hams must be stored at < 7 °C.

5) *Cooked Sausages.* Pieces of meat are cooked. After cooking they are partially cut or even minced, spices and salt are added. If liver is added the batter is thoroughly comminuted in a bowl chopper, filled in casings and reheated (*liver sausage*). Afterwards the products are kept under chilling conditions.
   For *blood sausages* the heated pieces of meat are mixed with blood, spices, and salt, filled in casings and reheated. Afterwards the products are kept under chilling conditions. *Aspic products* are obtained if the heated pieces of meat and often vegetables are inserted in dissolved hot gelatin (aspic) solutions with spices, acid and salt and heated in molds. They are stored under chilling conditions.
6) *Emulsion-Type Products*. Raw pieces of meat are comminuted with salt (1.5 – 3 % salt) in a bowl chopper, often additives and spices and about 25 – 35 % of water (usually crushed ice), finally fat meat or fatty tissue (e.g., pork back fat) are added. A further comminution follows until homogeneity is reached (6 to 9 min). The raw batter is filled in casings and heated to between 70 and 100 °C. The formed batter is sliceable at all temperatures. The products are stored at chilling temperatures.
7) *Meat Preparations.* Since the 1960s and 1970s a number of new meat preparations have appeared on the market. They include:
   – Products consisting of raw meat, covered with spices and salt. They are very perishable and must be treated as fresh meat.
   – Meat that is preheated and packed. If it is chilled it must be consumed within a few days.
   – Chilled meat that is minced or comminuted with salt and spices. These raw products should be treated as minced meat and consumed at the day of sale.

8. Manufacturing of Meat Products

8.1. Raw Meat Products

In principle, two kinds of raw meat products exist: Raw pieces of meat which are salted and dried (raw hams) and raw salami type sausages.

For all raw meat products meat with a low bacterial count is needed. This requires a good chilling regime after slaughter during the first 24 h post mortem to ≤ 7 °C and during dissecting of the carcass. Furthermore a low final pH between 5.4 and 5.8 is essential. Proper selection of meat during cutting and a subsequent freezing helps to achieve these conditions. Freezing does not only prevent the growth of microorganisms, by disintegration of cell walls (Section 6.2), it also permits that salt penetrates faster into the meat, lowering the water activity more quickly. Values of $a_w < 0.96$ are required which is reached at about 4.5 % salt in the meat. Fresh meat has a concentration of ions equivalent to 1 % NaCl and a_w values of 0.99.

8.1.1. Raw Hams

In order to prevent microbial growth the raw meat pieces are covered with crystalline salt or immersed in salt brines (8 – 20 % salt) at 3 – 5 °C [24]. The brine can be also injected with a multi-needle injector. By dry salting or immersion in brine the salt ions penetrate slowly into the meat. Depending on the size of the piece it takes several weeks until the NaCl concentration reaches the necessary 4 % NaCl in 1 – 2 cm depth. The temperature remains low until this salt concentration is reached. After taking off the brine the next phase is the equal distribution of salt. During this time the temperature rises slowly. The product is then partially cold smoked and dried at a relative humidity of 80 % and temperatures between 18 and 25 °C. After the drying process has reached the intended state, a further storage at 75 – 80 % relative humidity and about 10 °C is recommended to prevent further drying, growth of molds at the surface, and development of rancidity. The time before raw hams can be sold varies between two weeks for "fresh products" and two years for, e.g., Jamon Iberico.

Smoking of raw hams has nowadays primarily sensorial reasons. But smoking also prevents or at least retards growth of molds. In Mediterranean countries the traditional products have molds growing on the surface as they are not smoked. The molds produce flavor compounds which give Parma ham and others a part of their typical flavor.

Usually not only plain NaCl is added. It may contain nitrite (0.5 %) or nitrate. In some instances only pure salt is used, e.g., in Parma ham. The quality of the product shows that the processing does not necessarily need curing aids.

8.1.2. Raw Sausages

For raw sausages the same raw material requirements exist as for raw hams [24]. Low pH (< 5.8) and low bacterial load are essential. Again frozen meat is appropriate for this kind of sausages. Contrary to raw hams the material is ground or comminuted during the first step of processing. The grinding or comminution starts from meat frozen (− 5 °C) or very well chilled at − 1 °C. This low temperature is required in order to give the fat the necessary hardness. The cutting knives do not smear the fat, they cut through it. In the final product well separated particles of meat and fat are visible on slicing. To the ground material about 3 % curing salt or salt plus nitrate is added during further comminution. Spices, 0.1 – 1.5 % sugar (glucose, fructose, sucrose, lactose, maltose) are added, 50 ppm ascorbate and either glucono-δ-lactone (GdL) or starter cultures are mixed in. The inner anhydride GdL is rapidly hydrolyzed to gluconic acid, lowering the pH value of the mixture from 5.5 to about 4.9 within a few hours. The batter is more stable to many spoiling or pathogenic microorganisms. Starter cultures together with sugars also lower the pH value but act more slowly (in 2 – 3 days, pH < 5.3).

As starter cultures lactobacillae, micrococci, staphylococci, yeasts, or molds are used. The lactobacillae convert the sugars into lactic acid, which lowers the pH value. They themselves also retard or prevent the growth of spoiling or pathogenic microorganisms. Micrococae produce favorable flavor compounds from fats and proteins. Yeasts and molds support the flavor action of micrococci. Some yeasts also require oxygen for metabolism and thus prevent oxidative changes leading to rancidity.

The comminution ends at various degrees of size reduction, in practice reaching from roughly ground meat and fat particles down to a nearly homogeneous batter. End temperatures of − 2 to 4 °C are common.

The batters are filled in casings and subsequently fermented. The temperatures rise up to 25 °C in an atmosphere of 95 % relative humidity. The microorganisms grow and lower the pH. Slowly the relative humidity is reduced to 75 % and the temperature to 15 °C. Lowering of pH and an ongoing loss of water with a reduction of water activity to $a_w < 0.9$ take place within two to four weeks. The water content of the product is lowered from around 55 to 30 %, weight loss reaches up to 30 – 35 %. The salt concentrations increases from ca. 3 to above 4 %. Due to the water loss also protein and fat content increase. The product is ready for sale within two to four weeks. Packaged products may have shelf lives of several months at ambient temperatures, dependent on a_w, pH values, and packaging. Rather fresh and less dried products must be kept under chilling conditions (7 °C and lower).

In roughly ground products due to the drying process and the subsequent increase of salt concentrations, the myofibrillar proteins swell and form a gel at room temperature. The products are sliceable. Finely comminuted products do not form a gel so easily because this is prevented by fat layers between the protein chains. These products are spreadable.

The fat content of raw sausages varies from 30 to 50 %, depending on the amount of fat in the original product and the degree of drying.

8.2. Cooked Products

Cooked meat products constitute the main part of meat products. In central European countries about 65 % of all meat products consumed are heated products.

Three types of cooked products exist:

1) Emulsion-type sausages
2) Cooked (liver, blood) sausages and pate
3) Cooked ham

8.2.1. Emulsion-Type Sausages

Raw pieces of chilled meat are put in a bowl chopper [24]. In the rotating bowl a series of cutting knives comminute the meat. Salt, additives, and spices are added and about one half of the meat weight is added as crushed ice or cold water. The comminution continues until a

homogeneous batter is formed. The temperature is around 5 °C. Variable amounts of fatty tissue (e.g., back fat) are added and comminution continues until about 13 to 15 °C is reached when the batter is again homogeneous. The salt concentration at this stage varies between 1.8 and 3 % depending on the country and the product. The batter is filled in casings or cans and heated; in casings it may be smoked. The product is heated up to 75 °C for so-called "fresh sausages" and up to 120 °C in cans. By addition of salt myofibrillar proteins swell and a part even dissolves in the water. These proteins which have partially hydrophobic areas on their surface surround small fat particles (10 – 100 µm). On heating the proteins form a three-dimensional network, a heat-stable gel. The sausages become sliceable. If nitrite curing salt is added, the nitrite reacts with myoglobin forming nitrosomyoglobin which has an heat-stable red color.

The finely comminuted batter may be mixed with precured pieces of meat, vegetables (mushrooms), or nuts and filled in casings and heated. Usually the fat contents of the finely comminuted batter is around 20 to 25 %.

8.2.2. Cooked Sausages

This group consists of liver and blood sausages, pates, and aspic products.

For *liver sausages* the meat and fat are precooked. The heated material is comminuted, part of the cookout is taken off, the material cools below 60 °C and raw liver is comminuted with spices, herbs and salt and added. The mixture is cooled down to about 35 °C, filled in casings or cans, and reheated up to 75 – 110 °C. Sausages in casings may be smoked during the second heating. The degree of comminution may vary. Liver sausage is a spreadable product as no heat-stable network is formed. Salt is added to achieve hygienic stability and for sensorial reasons. There is no technological need for salt addition as in emulsion-type sausages. Pate production follows the same pattern.

For *blood sausages* the cooked — often precured — meat and fat are cut or coarsely ground, raw blood, spices, salt, and often gelatin are added, the mixture filled in casings and reheated to 75 – 110 °C.

With *aspic products* the same procedure is used. Cooked meat pieces are mixed with spices, salt, organic acids (mainly vinegar), often also with vegetables, and placed in molds with a gelatin solution. The mold is reheated again to 75 – 95 °C.

Neither liver sausages, pates, blood sausages nor aspic products form an heat-stable gel. They may become sliceable in the chilled state.

8.3. Cooked Ham

Whole deboned cuts of hind legs of pigs are brined by injection, smaller pieces of lean meat from all parts of a carcass are put in a so-called tumbler with salt brine and spices. The meat pieces in the tumbler are slowly moved at chill room temperature. During this process the salt penetrates into the myofibrillar structures and lets the proteins swell and partially dissolve. After about 24 h the material is put in casings or molds and heated to 65 – 75 °C. The proteins form a heat-stable network. The cooked ham is sliceable. A traditional product has a yield of 100 – 110 %. Addition of additives which bind water (e.g., nonmeat proteins, agar and other carbohydrate polymers) permit yields of well above 110 %. Some countries call these products *ham and water*. If only lean meat is used, the fat content is below 5 %. Ham and water products may contain less then 2 % fat (more than 98 % fat free).

9. Meat Species Identification

The meat species identification is of general interest for trade and consumers alike. Both want to know which meat species has been used for a meat meal or which species are mixed in a meat product. In various religious and cultural societies some species are not permitted, e.g., pork in Jewish or Moslemic societies, beef in Hinduistic countries. With the upcoming of BSE also the consumers in Europe started to reject beef.

Species identification can be done via detection of

- Proteins specific for the species,
- Fatty acid composition of fats, or
- DNA analysis [25].

Detection of specific proteins can be done by electrophoretic or immunological methods.

Electrophoresis. In electrophoresis polyacrylamide gels (PAGE) proteins are used either in the native or the denatured form (sodium dodecyl sulfate polyacrylamide gel, SDS – PAGE). Denaturation is done with SDS which is absorbed on the surface of the denatured proteins, leading to negative charges on the protein globule due to the sulfate groups. The separation in the SDS – PAGE gel depends on the molecular mass of the proteins. PAGE separates according to the electric charges and the size of the molecule. Proteins may be positively or negatively charged.

In isoelectric focussing (IEF), the proteins are fractionated according to their isoelectric point, by migration along a continuous pH gradient.

All electrophoretic methods work with raw meat extracts or mildly heated meat (< 110 °C). At higher temperatures an intensive denaturation or cleaving of the amino acids chains limit the electrophoretic methods.

Immunological methods with an antigen-antibody interaction open a wider field. Specific antibodies for each species and different denaturation structures permit to detect concentrations of 1 – 2 % of a species in mixtures. Modern enzyme-linked immunosorbent assay (ELISA) tests permit the detection in products heated above 120 °C.

Fatty Acid Analysis. The fatty acid analysis is the most uncertain of all methods. In monogastric species like pigs or poultry the fatty acid composition of fats depends on the fats in the feed. In polygastric species such as cattle or sheep the influence of feed is smaller but still existing. Therefore the source of the material must be known as a prerequisite for the species identification via the fat composition. A species identification on rare fatty acids with more than 20 C-atoms in the chain has been reported [26].

DNA Analysis. In the DNA analysis specific segments of base pairs of a gene can be amplified by polymerase chain reaction (PCR) (→ Polymerase Chain Reaction). A number of species specific primers exist on the market. The PCR test is rather sensitive but not necessarily quantitative. False positive reactions occur already with traces of a species in a mixture. The PCR tests work satisfactorily even above 120 °C up to 130 °C.

10. World Meat Production

Worldwide well above 260×10^6 t of meat were produced in 2004 [27]. Amount of production means produced carcass weight. That is the dressed carcass without eviscera and genitals, without some fat parts and, if customary, without skin but with bones, because the carcasses are usually weighed on the scale at the end of the slaughter process.

The main species used for meat production are pigs. Pork is produced with ca. 101×10^6 t/a. It is followed by poultry (sum of mainly chicken and turkey but also goose and duck) with 77×10^6 t/a. Beef is produced in slightly smaller quantities of around 62×10^6 t/a. Sheep plus goat meat follow with about 12×10^6 t/a [27].

The main *pork* producing area is Asia, especially East Asia, with about 50 % of the worldwide production, followed by Europe with 25 % of the world production. North America contributes 12 % of the annual pork production. Asia also leads in *poultry* production with about one third of the world annual production. North America follows with 25 % annually. Europe produces 17 % of the annual world production. *Beef* production is led by North America with around 27 % of the world production. South America, Asia, and Europe follow each producing between 15 and 19 %. Asia is again the main producer of sheep plus goat meat with about 45 % of the world production. Africa follows with 30 %, Europe with 27 %, and Oceania with 11 %.

All other meat sources are unimportant in comparison. Around 5×10^6 t/a go to all other meats including horse meat, and game.

For many years the meat production has increased annually by 2 – 4 % worldwide. The lowest growth rate is observed in Europe with a surplus meat production. Despite the same economic situation in North America its meat production is slowly growing but more quickly than in Europe. The strongest increase is found in Asia.

11. Meat Grading

One aspect to value a carcass is its dressed weight. The dressed weight consists of lean meat, fat, and bones. The edible parts of a carcass consist of muscular and fatty tissues in various relationships and locations and of some edible offals. The offals vary little in percentage of carcass weight.

The most expensive part of the carcass is the lean meat (muscular tissue). Fat is much less esteemed and bones are cheap byproducts. But the fat in a carcass is often not well separable from the muscular tissue. Muscular tissue may contain thin threads or sheets of fatty tissue. Then the meat is called marbled. The meat content of a carcass must be estimated in order to pay the producer according to the weight and meat content; also customers will buy selected carcasses or carcass parts according to the meat content in order to achieve an optimal use of the meat. Also the proportions of carcass parts are of importance. Large hams of otherwise lean carcasses are welcomed by ham producers.

Due to the unequal distribution of lean meat and fat in a carcass no direct measurement methods of the meat and fat content exist. Until 1980 the meat content was judged by trained graders. This is still the case with cattle carcasses. For pig carcasses instruments were developed in the 1970s which, at certain positions in the back of a carcass, measure, for instance the thickness of back fat and muscular tissue diameter of the loin muscle [28]. The probes insert a needle into the carcass, the tip of the needle having two sets of glass fibers build in. One set of glass fiber emits light, the other set measures the reflection. Light fatty tissue reflects differently than red meat. During insertion a measurement is taken every 0.5 mm and a graph of reflection with the distance is obtained (Fig. 6). An algorithm calculates the meat content of a carcass from the obtained fat and muscles thickness.

Other grading systems measure length, width, and thickness of a carcass part. Furthermore angles (convex and concave forms) of some parts such as hind legs are measured. This is best done with video cameras in three dimensions. A practicable system is being developed at the moment.

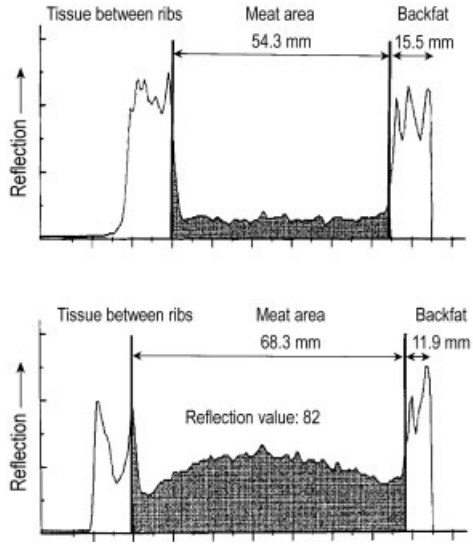

Figure 6. Diagram of the pork grading system Fat-o-Meat'er. A measurement value is taken every 0.5 mm. On the right-hand side the reflection of white backfat is depicted, in the center the reflection of the meat in the loin muscle, and on the left-hand side the reflection of fatty and connective tissue between ribs. In the upper picture the reflection of a red, normal muscle is shown; the lower diagram exhibits the higher reflection of pale soft exudative (PSE) meat (courtesy of BRANSCHEID [28]).

Other methods use ultrasound scanning. Ultrasound is reflected at the border of different tissues (fat and muscles). Also the sound velocity is different in different tissue. A number of instruments detect either reflection, velocity of penetration, or both.

X-ray and NMR tomography are also available. In the last years total body electric conductivity (TOBEC) or bioelectric impedance (BIA) have been recommended for meat grading.

All methods of measurement used are indirect methods which are connected with estimation errors. The errors, however, are usually small.

12. References

1. EU-Directive 64/433/EWG in its amended version. Official Journal of the European Communities, L 268, 71, 24.9.1991 (1991)
2. E. D. Aberle, J. C. Forrest, P. E. Gerrard, E. W. Mills: *Principles of Meat Science*, Kendall/Hunt, Publishing Company Dubuque, Iowa, 2001.

3. *Lawrie's Meat Science*, 6th Edition, Woodhead, Cambridge, 1998.
4. A. M. Pearson, R. B. Young: *Muscle and Meat Biochemistry*, Academic Press, Inc. San Diego, New York 1989.
5. H. Kofrany, S. Wirths: *Einführung in die Ernährungslehre*, Umschau Verlag, Frankfurt/Main 1987.
6. J. Seuss-Baum: "Ernährungsphysiologische Bedeutung von Fleisch und Fleischerzeugnissen," in W. Branscheid, K. O. Honikel, G. von Lengerken, K. Troeger (eds.): *Qualität von Fleisch und Fleischwaren*, Deutscher Fachverlag, Frankfurt/Main 1998, p. 617.
7. S. W. Souci, W. Fachmann, H. Kraut (eds.): *Die Zusammensetzung der Lebensmittel*, 5th ed., Medpharm, Scientific Publishers, Stuttgart 1994.
8. E. Bjorn-Rasmussen, L. Halberg: "Effect of animal proteins on the absorption of food iron in man," *Nutr. Metab.* **23** (1979) 129.
9. EU-Directive for Meat Hygiene 95/23/EG, Official Journal of the European Communities, L 243, 11.10.1995 (1995)
10. K. Ender, C. Augustini: "Schlachtwert von Rind und Kalb," in W. Branscheid, K. O. Honikel, G. von Lengerken, K. Troeger (eds.): *Qualität von Fleisch und Fleischwaren*, Deutscher Fachverlag, Frankfurt/Main 1998, p. 165.
11. K. O. Honikel, F. Schwägele: "Biochemische Prozesse der Fleischbildung in: Qualität von Fleisch und Fleischwaren," in W. Branscheid, K. O. Honikel, G. von Lengerken, K. Troeger (eds.): *Qualität von Fleisch und Fleischwaren*, Deutscher Fachverlag, Frankfurt/Main 1998, p. 593.
12. E. Dransfield: "Modelling post-mortem tenderization IV. Role of calpains and calpastatin in conditioning," *Meat Science* **34** (1995) 217.
13. C. L. Davey, K. V. Gilbert: "The temperature coefficient of meat ageing," *J. Sci. Food Agric.* **27** (1976) 244.
14. S. L. Beilken, P. E. Bouton, P. V. Harris: "Some effects on the Mechanical Properties of Meat, Produced by cooking at Temperatures between 50° and 60 °C," *J. Food Sci.* **51** (1986) 791.
15. F. K. Lücke, K. Troeger: "Fleischhygiene: Mikrobiologische Risiken," in W. Branscheid, K. O. Honikel, G. von Lengerken, K. Troeger (eds.): *Qualität von Fleisch und Fleischwaren*, Deutscher Fachverlag, Frankfurt/Main 1998, p. 439.
16. R. Lautenschläger: "Diffusionsverhalten der Salze und Wertung der Technologien beim Pökeln." Dissertation, Technical University, Berlin 1995.
17. "Nitrat, Nitrit, Nitrosamine in Lebensmitteln — Statusbericht 1995," E. Mayer-Miebach (ed.): *Berichte der Bundesforschungsanstalt für Ernährung*, Karlsruhe, Germany 1995.
18. EU-Directive 95/2/EG, Official Journal of the European Communities, L 61, 1, 18.3.1995 (1995)
19. EU-Directive 94/36/EC, Official Journal of the European Commission, L 237, 13, 10.09.1994 (1994).
20. L. Toth (ed.): *Chemie der Räucherung*, Verlag Chemie, Weinheim, Germany 1982.
21. EU Regulation EC, No. 208/2005, Official Journal of the European Commission L 34, 3, 08.02.2005 (2005).
22. Food Irradiation. A technique for preserving and improving the safety of food, WHO, Geneva 1988.
23. J. C. Cheftel: "Effects of High Pressure on Meat: A Review," *Meat Science* **46** (1997) 211.
24. K. Troeger, A. Fischer, C. Hilmers, R. Lautenschläger, W. D. Müller, A. Stiebing: "Fleischwaren," in W. Branscheid, K. O. Honikel, G. von Lengerken, K. Troeger (eds.): *Qualität von Fleisch und Fleischwaren*, Deutscher Fachverlag, Frankfurt/Main 1998, p. 773.
25. K. Hofmann: "Tierartbestimmung bei Fleisch und Fleischwaren," in W. Branscheid, K. O. Honikel, G. von Lengerken, K. Troeger (eds.): *Qualität von Fleisch und Fleischwaren*, Deutscher Fachverlag, Frankfurt/Main 1998, p. 678.
26. E. Lücker, W. Biedermann, S. Lachhab, U. Truyen, A. Hensel: "GS-MS-detection of central nervous tissues as TSE risk material in meat products: analytical quality and strategy," *Anal Bioanal. Chem.* **380** (2004) 866.
27. FAO-Food outlook No. 4, Rome, December 2004. www.fao.org/docrep/007/j3877e/j3877e08.tm.
28. W. Branscheid, G. von Lengerken: "Die Erfassung der Schlachtkörperzusammensetzung und die Einstufung in Handelsklassen," in W. Branscheid, K. O. Honikel, G. von Lengerken, K. Troeger (eds.): *Qualität von Fleisch und Fleischwaren*, Deutscher Fachverlag, Frankfurt/Main 1998, p. 97.

pH Measurement and Control

HELMUTH GALSTER, Mettler-Toledo GmbH, Process Analytics, Bad Vilbel, Federal Republic of Germany

1. Introduction . . . 2
1.1. Historical Background . . . 2
1.2. Buffer Solutions . . . 2
1.3. Activities . . . 3
1.4. Self-Dissociation . . . 3
1.5. Importance of pH Measurements . . . 4
2. pH Scales . . . 5
2.1. Indicators . . . 5
2.2. Electrochemical Methods . . . 5
2.3. Primary pH Standards . . . 6
2.4. Secondary pH Standard Solutions . . . 6
3. Electrode Cells . . . 8
3.1. Glass Electrodes . . . 8
3.2. Reference Electrodes . . . 11
3.3. Liquid Junctions . . . 12
3.4. Combined Electrodes . . . 13
3.5. Testing Glass Electrode Cells . . . 14
3.6. Alternative pH Electrodes . . . 16
4. pH Meters or Transformers . . . 16
4.1. Input Amplifier . . . 16
4.2. Correction Functions . . . 16
4.3. Output . . . 17
4.4. Microprocessor-Controlled pH Meters . . . 17
4.5. Testing pH Meters . . . 18
4.6. Self Monitoring . . . 18
4.7. Self Calibration . . . 18
5. Analytical Applications . . . 19
5.1. Calibration of Glass Electrode Cells . . . 19
5.2. pH Measurement in Water Analysis . . . 20
5.3. pH Measurement in Concentrated Solutions . . . 20
5.4. Suspension Effects . . . 21
5.5. pH Measurement in Nonaqueous Solvents . . . 21
6. Technical pH Measurements . . . 22
6.1. Extreme Conditions . . . 22
6.2. Installation . . . 23
6.3. Electrode Assemblies . . . 23
6.4. Pressure Control . . . 24
6.5. Retractable Assemblies . . . 25
7. pH Control . . . 25
8. References . . . 28

1. Introduction

1.1. Historical Background

In the 1500s THURNEYSSER [1] discovered that violet sap was colored different hues by sulfurous and sulfuric acids. This first indicator was employed over a long period to detect acids. Hydrogen and hydroxide ions acquired their special importance as the dissociation products of water when ARRHENIUS [2] and OSTWALD [3] put forward the ionic theory in the 1880s.

In 1904 FRIEDENTHAL [4] set up the first scale for classifying acids by determining the dissociation constants of weak acids by conductivity measurements and using 14 indicator dyes with color changes at particular hydrogen ion concentrations. He also used this concept for classification of bases. When it became apparent that the dilution of acids or bases did not yield reliable solutions with low hydrogen ion or hydroxide ion concentrations, FRIEDENTHAL, on the advice of PAL SZILY, used the mixtures of salts and acids, that are now referred to as buffer solutions.

FRIEDENTHAL's hydrogen ion concentrations of 10^{-14} to 1 mol/L were awkward to handle. To simplify this problem SØRENSEN [5] suggested taking the negative logarithm of the hydrogen ion concentration. He named it the "hydrogen exponent" and introduced the term pH (*pondus hydrogenii*) for it.

1.2. Buffer Solutions

Today pH values are represented by a series of standard solutions. These standards must maintain their pH values even when contamined with small quantities of acids or bases or when diluted with a little water. *Buffer Solutions* possess this property. Typical buffer solutions consist of mixtures of a weak acid and a salt of this acid with a strong base, or of a weak base and a salt of this base with a strong acid [6].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.e19_e01

According to the law of mass action (LMA) and ignoring activity coefficients (Section 1.3) it follows for pure acetic acid (HA), for example:

$$c\left(\mathrm{H}^{+}\right) c\left(\mathrm{A}^{-}\right) \approx c(\mathrm{HA}) K_{\mathrm{a}}$$
$$K_{\mathrm{a}} = 10^{-4.74} \mathrm{mol/L} \tag{1.1}$$

If sodium acetate (NaAc) is added, the concentration of acetate ions is equal to that of the added salt C(NaAc) plus that of the dissociated portion of the acetic acid:

$$c\left(\mathrm{A}^{-}\right) = C(\mathrm{NaAc}) + c\left(\mathrm{H}^{+}\right) \tag{1.2}$$

Since $c(\mathrm{H}^{+}) \ll C(\mathrm{NaAc})$, the hydrogen ion concentration can be neglected. Hence

$$c\left(\mathrm{A}^{-}\right) \approx C(\mathrm{NaAc}) \tag{1.3}$$

and since the amount of dissociated acid is low, the actual concentration is approximately equal to the initial concentration C(HA)

$$c(\mathrm{HA}) = C(\mathrm{HA}) \tag{1.4}$$

Then we obtain

$$C\left(\mathrm{H}^{+}\right) C(\mathrm{NaAc}) \approx C(\mathrm{HA}) K_{\mathrm{a}} \textit{ or}$$
$$-\log C\left(\mathrm{H}^{+}\right) \approx \mathrm{pH} \approx -\log K_{\mathrm{a}}$$
$$+\log \frac{C(\mathrm{NaAc})}{C(\mathrm{HA})} \tag{1.5}$$

Equations (1.5) were derived by using approximations that neglect the activity coefficients. This pH is therefore only suitable for approximate calculations. The true pH must be determined experimentally.

The pH values of some buffer solutions are determined exactly and serve as standards for pH measurements (see Section 2.3).

Quantitative reactions of the pH of a buffer solution to external influences are measured in terms of the buffer capacity. The *buffer capacity* is the ration of the differential amount of added substance to the pH change that it causes.

When the additive is a strong acid or base and is measured in gram equivalents n relative to the buffer volume V_{o}, then

$$\beta = \frac{1}{V_0} \left| \frac{\mathrm{d}n}{\mathrm{d}(\mathrm{pH})} \right| \tag{1.6}$$

where β is the *buffer value* in $\mathrm{mol\,L^{-1}\,(pH)^{-1}}$. The second factor in Equation (1.6) is the reciprocal gradient of the titration curve as the titration is commenced. The buffer value is important not only for buffer solutions but also for estimating the ability to control the pH of process solutions. For practical determination of buffer values, see [7].

1.3. Activities

The chemical effectiveness of ions depends not only on their concentration but also on interactions between them. To allow for these deviations it is necessary to multiply the concentrations b [mol/kg] of each substance by its *activity coefficient* γ. The product of concentration and activity coefficient is the *activity* $a = b\,\gamma$. The *law of mass action* (LMA) describes the dissociation of an acid HA correctly as follows:

$$\frac{b\left(\mathrm{A}^{-}\right) \gamma\left(\mathrm{A}^{-}\right) b\left(\mathrm{H}^{+}\right) \gamma\left(\mathrm{H}^{+}\right)}{b(\mathrm{HA})\, \gamma(\mathrm{HA})} = \frac{a\left(\mathrm{A}^{-}\right) a\left(\mathrm{H}^{+}\right)}{a(\mathrm{HA})} \tag{1.7}$$

The correct definition of pH is

$$\mathrm{pH} = -\log \frac{b\left(\mathrm{H}^{+}\right) \gamma\left(\mathrm{H}^{+}\right)}{b^{0}} = -\log \frac{a\left(\mathrm{H}^{+}\right)}{b^{0}} \tag{1.8}$$

where $b^0 = 1$ mol/kg (the units of a).

The condition of electrical neutrality requires that cations and anions are always simultaneously present. Therefore, individual ion activities do not represent any physical reality, and only a geometric *mean activity coefficient* $\gamma_{\pm}$ can be determined directly. For example, the coefficient for sodium chloride is

$$\gamma_{\pm} = \sqrt{\gamma\left(\mathrm{Na}^{+}\right) \gamma\left(\mathrm{Cl}^{-}\right)} \tag{1.9}$$

A large number of mean activity coefficients are tabulated in [8]. They are determined from thermodynamic measurements such as depression of freezing point. In each case the activity coefficient approaches unity in dilute solutions as $b \to 0$.

To obtain an indication of the magnitude of the activities of individual ions, it is necessary to calculate them by theoretical means. The DEBYE–HÜCKEL *theory* of *interionic interaction* [9], developed in 1923, is applicable for ionic strengths in the important range up to 0.1 mol/L. For an ion i

$$\log\gamma = -\frac{Az_i^2 I^{1/2}}{1 + B\text{Å}I^{1/2}} \tag{1.10}$$

where z_i is the charge number and Å is the ionic diameter in Ångstrom. A and B are constants which are only dependent on the relative dielectric constant ε and the temperature T; in water at 25 °C, $A = 0.5108$ and $B = 0.3286$. Values of A and B at other temperatures are tabulated in [10], [11]. However, it is not possible to calculate the diameter of the hydrated ions *Å* accurately; it can only be estimated approximately from crystal data, viscosities, ionic mobilities, etc. The best results available at present are given in [12].

In Equation (1.10) I represents the *ionic strength* expressed in terms of amount-of-substance concentration and is defined as half the sum of the concentrations of all ionic species in solution multiplied by the squares of their charges:

$$I_\mathrm{b} = \frac{1}{2}\sum\left(b_i z_i^2\right) \tag{1.11}$$

At ionic strengths above 1 mol/L the logarithm of the activity coefficient becomes positive, and an additional term is required in Equation (1.10), [8], [13].

The term *ÅB* in Equation (1.10) introduces the greatest amount of uncertainty. According to a convention adoped by BATES and GUGGENHEIM [14] in the case of chloride ions the product *ÅB* is assumed to have the value 1.5 at all temperatures; this is accepted at present as the most likely value. Thus, the DEBYE – HÜCKEL equation becomes

$$\log\gamma\left(\mathrm{Cl}^-\right) = -\frac{AI^{1/2}}{1 + 1.5I^{1/2}} \tag{1.12}$$

It allows the comparison of measured activities.

1.4. Self-Dissociation

The self-dissociation of water is generally written in the simplified form:

$$\mathrm{H_2O} \rightleftharpoons \mathrm{H^+} + \mathrm{OH} \tag{1.13}$$

However, hydrogen ions cannot exist as free protons in any solution. Since it is an elementary particle with a very high charge density, the proton exerts large forces on the surrounding solvent molecules, so that strong electrostatic bonds with the solvent molecule are produced. The first stage in the case of water is the formation of the *oxonium ion*:

$$\mathrm{H^+} + \mathrm{H_2O} \rightleftharpoons \mathrm{H_3O^+} \tag{1.14}$$

The self-dissociation or *autoprotolysis* of water is correctly written as

$$2\,\mathrm{H_2O} \rightleftharpoons \mathrm{H_3O^+} + \mathrm{OH^-} \tag{1.15}$$

The reaction occurs without any measurable treshold and is only limited by the collision frequency [15]. The dissociation constant was first determined by measuring electrical conductivities [16]. The *autoprotolysis constant* of water at 25 °C is:

$$K_\mathrm{aq} = \frac{a\left(\mathrm{H_3O^+}\right)a\left(\mathrm{OH^-}\right)}{a^2\left(\mathrm{H_2O}\right)} = 10^{-15.742}\,\mathrm{mol/L} \tag{1.16}$$

Since the dissociation is only slight, the activity of water [$\log a(\mathrm{H_2O}) = 1.745$] is essentially unchanging; hence, it can be incorporated into the constant. The *ionic product* of water at 25 °C is therefore given in the simplified form by:

$$K_\mathrm{w} = a\left(\mathrm{H^+}\right)a\left(\mathrm{OH^-}\right) = 10^{-13.997}\,(\mathrm{mol/L})^2 \tag{1.17}$$

The consequence is a pH value for pure water of 7.0, and a solution of a strong hydroxide with $a = 1$ mol/L has a pH of 14.

Nonaqueous, but water-like, solvents also self-dissociate in a manner analogous to water. According to the general equation, for HSol as solvent, it follows that

$$2\,\mathrm{HSol} \rightleftharpoons \mathrm{H_2Sol^+} + \mathrm{Sol^-} \tag{1.18}$$

with the autoprotolysis constant

$$K_\mathrm{Sol} = \frac{a\left(\mathrm{H_2Sol^+}\right)a\left(\mathrm{Sol^-}\right)}{a^2\left(\mathrm{HSol}\right)} \tag{1.19}$$

The constant K_Sol depends on the permissivity of the solvent, which ranges from $\varepsilon = 2$ for benzene to $\varepsilon = 182$ for *N*-methylacetamide (water: $\varepsilon = 78$). The range of pH scale in individual solvents is equal to the value of $\mathrm{p}K_\mathrm{Sol}$; this is illustrated graphically in Figure 1. The halfway

point of the scale, where cation and anion concentrations of the solvent are equal, is the neutral point on the pH scale in the solvent concerned. For detailed tables, see [18].

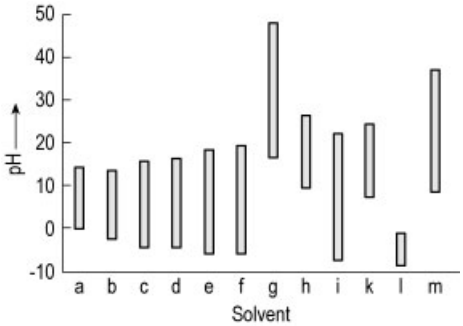

Figure 1. Relative positions of pH scales in different solvents [17]
a) Water; b) Methanol; c) Ethanol; d) 1-Propanol; e) 1-Butanol; f) Isobutyl alcohol; g) Ammonia; h) Hydrazine; i) Acetonitrile; k) Formamide;) l) Formic acid; m) Dimethyl sulfoxide

To compare activities in nonaqueous solutions with those in aqueous solutions we must distinguish between primary and secondary *medium effects*:

$$\log {}^{\mathrm{s}}_{\mathrm{w}}\gamma_i = \log {}^{\mathrm{s}}_{\mathrm{w}}\gamma_i^0 + \log {}^{\mathrm{s}}_{\mathrm{s}}\gamma_i \tag{1.20}$$

where ${}^{\mathrm{s}}_{\mathrm{w}}\gamma_{\mathrm{i}}$ is known as the total medium effect, and ${}^{\mathrm{s}}_{\mathrm{w}}\gamma_{\mathrm{i}}^0$ [19] as the *primary medium effect*. The latter incorporates the activity change which is associated with the change of medium from water to solvent. ${}^{\mathrm{s}}_{\mathrm{s}}\gamma_{\mathrm{i}}^0$ is the *secondary medium effect* or concentration effect of substance *i*, which corresponds to the activity coefficient described above.

1.5. Importance of pH Measurements

The knowledge of the pH value is important in many areas:

- Determination of natural constants, e.g., acid constants
- Chemical industries need pH control in all production stages
- Biotechnology: bacterial growth depends on exact pH control
- One of the first applications was pH control in testing of foodstuffs
- In medicine, pH measurings are part of many diagnoses
- Environmental technology uses pH measurements for monitoring rain and surface waters; pH is the most important parameter.
- Wastewater cannot be regenerated without controlling pH
- The preparation of drinking water is impossible without a pH control

Table 1 lists approximate pH values of some everyday solutions.

Table 1. pH values of everyday solutions [105]

Solution	pH value	Solution	pH value
Gastric juice	ca. 1	Saliva	ca. 6.7
Lemon juice	2.3	Pure water	7.00
Vinegar	3.1	Blood	7.36
Beer	4.3	Sea water	8.3
Sour milk	4.4	Soap suds	8.4
Plum juice	5 – 6	Lime water	12.3
Fresh milk	6.5	Sodium hydroxide $c = (1\ \mathrm{mol/L})$	ca. 14

2. pH Scales

2.1. Indicators

The alteration of the color of dyes under the influence of acids or bases is one of oldest observations in chemistry. The first pH scale, which was set up by FRIEDENTHAL [4], consisted of a table of indicator dyes. For a comprehensive review of modern indicators, see → Indicator Reagents.

2.2. Electrochemical Methods

Practical pH. In industrial chemistry exclusively electrochemical pH measurements are made. A potentiometric cell is used to make a comparison with *standard solutions*. For a single standard [20] we have:

$$\mathrm{pH} = \mathrm{pH\,(S)} - \frac{U - U\,(S)}{k} \text{ with } k = \frac{RT\ln 10}{F} \tag{2.1}$$

where pH(S) is the pH value of the standard solution, U(S) the corresponding voltage of the cell and k the slope of the cell according to the NERNST Law [21]. Values for several temperatures are listed in Table 2.

Table 2. Temperature dependence of the Nernst voltage $k = (RT/F)\ln 10$ [22]

t, °C	k	t, °C	k	t, °C	k
0	54.199	35	61.144	70	68.088
5	55.191	40	62.136	75	69.081
10	56.183	45	63.128	80	70.073
15	57.175	50	64.120	85	71.065
20	58.167	55	65.112	90	72.057
25	59.159	60	66.104	95	73.049
30	60.152	65	67.096		

The Hydrogen Electrode. In 1893 LEBLANC [23] discovered that some metals behave *like hydrogen electrodes* when hydrogen is bubbled over them. Later it was found that platinum metals with hydrogen act as ideal selective electrodes for hydrogen ions. The simple electrode reaction on a platinum metal is

$$H_2 \rightleftharpoons 2\,H^+ + 2\,e^- \tag{2.2}$$

The galvanic potential of the electrode follows the NERNST Law exactly:

$$\varphi_{H_2} = \frac{RT}{F}\ln a\left(H^+\right) - \frac{RT}{2F}\ln P\left(H_2\right) \tag{2.3}$$

For all measurements with the hydrogen electrode the hydrogen pressure must be corrected [24].

When $a(H^+) = 1$ mol/kg and $P(H_2) = 1$ atm, the electrode is called the *standard hydrogen electrode*. Its potential is defined as zero at all temperatures. All potentials relative to this electrode are called *standard potentials*.

The Harned Cell. To obtain exact measurements of hydrogen ion activities, a cell with the hydrogen electrode and a *reference electrode* is required. A silver/silver chloride electrode and an electrolyte containing chloride ions is used. Without a liquid junction, this cell is known as a *Harned cell* (Figure 2). The cell

$$\mathrm{Ag}\,\Big|\,\mathrm{AgCl, HCl}\,\Big|\,\mathrm{H_2, Pt} \tag{2.4}$$

produces a measurable voltage when $P(H_2) = 1$ atm:

$$U = -U^0_{\mathrm{AgCl}} + k\log a\left(Cl^-\right) + k\log a\left(H^+\right) \text{ or} \tag{2.5}$$

$$U = -U^0_{\mathrm{AgCl}} + k\left[\log b\left(H^+\right) + \log\gamma\left(H^+\right) + \log b\left(Cl^-\right) + \log\gamma\left(Cl^-\right)\right] \tag{2.6}$$

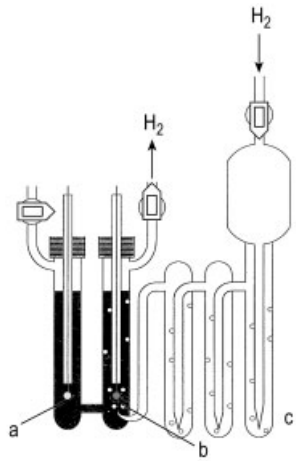

Figure 2. Harned cell
a) Ag/AgCl electrode; b) Pt/H_2 electrode; c) Moistening of hydrogen
Courtesy of PTB, Braunschweig

The Ag/AgCl electrode is one of the most widely reference electrodes. It consists of silver metal which is covered with silver chloride. The standard potential U^0_{AgCl} is determined in water [25] and in water – methanol mixtures [26].

2.3. Primary pH Standards

Extrapolation. If the concentration $b(Cl^-)$ [mol/kg] is known, the influence of $\gamma(Cl^-)$ can be estimated. The only certain value is $\gamma(Cl^-) = 1$ at the concentration $b(Cl^-) = 0$. This point is arrived at by extrapolation with stepwise decreasing chloride concentrations. The chloride ion activity must be calculated with the DEBYE – HÜCKEL equation and the BATES – GUGGENHEIM convention, (see Section 1.3) [27], [28]. This calculation allows an extrapolation with a straight line. The result is shown graphically in Figure 3 for a phosphate buffer as an example.

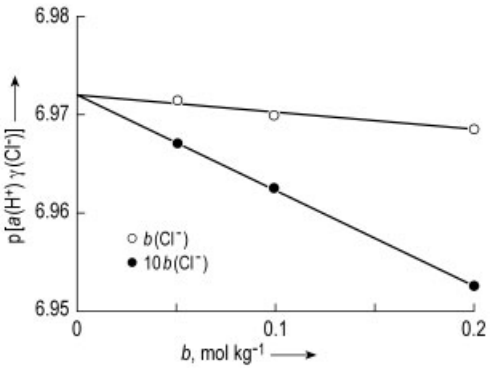

Figure 3. Dependence of p[$a(H^+)\gamma(Cl^-)$] for 0.025 mol/kg phosphate buffer solution on $b(Cl^-)$ [28]

A buffer solution measured in this manner represents at present the best approximation to the hydrogen ion activity defined at

$$\mathrm{pH} = -\log \frac{a\left(\mathrm{H}^+\right)}{m^0} \tag{2.7}$$

Reference Buffers. Some national institutes, (e.g., the *U.S. National Institution of Standards* (NIST) in Washington; the *Physikalisch-Technische Bundesanstalt* (PTB), in Braunschweig, Germany), carry out reference measurements. Substances form tested chemical batches are *certified* and brought on the market as *reference buffer materials* by *accredited* manufacturers [29].

These chemicals are dissolved to form *reference buffer solutions*. Table 3 lists concentrations and pH values of nine *primary standard buffer* solutions. Differences in pH of ± 0.005 between different batches are possible. These solutions represent a pH scale (multipoint scale, DIN 19266) [30].

2.4. Secondary pH Standard Solutions

Differential Potentiometric Cell. The primary pH reference solutions are too expensive for practical use. Commercial buffer solutions are therefore produced by comparison with primary reference buffer solutions. To compare a commercial buffer solution with a primary reference solution, a symmetric cell with two hydrogen electrodes and a liquid junction is used (Figure 4).

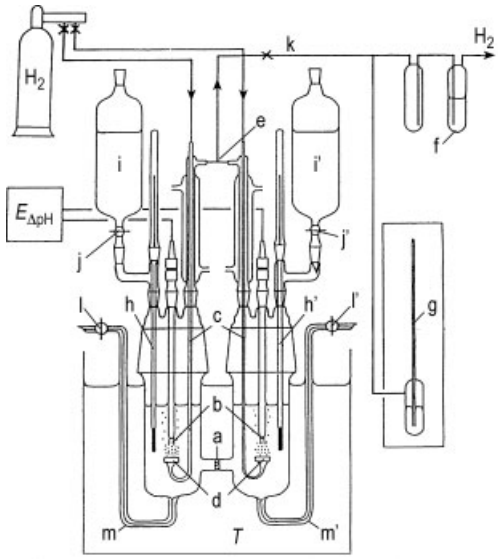

Figure 4. Differential potentiometric cell with identical Pt, H_2 electrodes [31]
a) Sintered disk; b) Platinum electrodes; c) Glass tubes; d) Porous glass disks; e) T-tube; f) Bubble flasks; g) Manometer tubes; h) Thermometers; i) Cylindrical funnels for buffer solutions; j, k, l) Stopcocks; m) Outlet tubes

Since the compared solutions are quasi-identical, the liquid junction potential can be neglected. Extensive tests verify that the additional uncertainty is not more than $\Delta\mathrm{pH} = \pm 0.003$ [31]. The two hydrogen electrodes must be exactly isothermal. Measurements should be made at more than one temperature. The uncertainty of secondary reference pH buffer solutions is $\Delta\mathrm{pH} \approx \pm 0.004$.

Technical Buffer Solutions. The primary buffers are relatively dilute solutions with only small buffer values. For technical use higher concentrations with buffer values $\beta > 0.05$ are required. These technical buffers are compared with primary reference buffer solutions, for which the differential cell is not suitable. The measurements are performed with a conventional glass electrode cell (Chap. 3). It is important to use two standard solutions which include the technical buffer. Table 4 lists some examples of buffer solutions recommended by DIN 19267 [7].

Table 3. Examples of primary puffer solutions∗ [30]

	A	B	H	C	D	E	F	I	G
t °C									
5	1.668		3.837	4.004	6.950	7.502	9.392	10.248	13.207
10	1.670		3.819	4.001	6.922	7.474	9.331	10.180	13.003
15	1.672		3.801	4.001	6.900	7.451	9.277	10.121	12.810
20	1.676		3.787	4.003	6.880	7.432	9.228	10.066	12.627
25	1.680	3.557	3.775	4.008	6.865	7.416	9.184	10.014	12.454
30	1.680	3.553	3.766	4.015	6.853	7.405	9.144	9.970	12.289
37	1.694	3.550	3.757	4.028	6.841	7.393	9.095	9.911	12.043
40	1.697	3.549	3.754	4.036	6.837	7.389	9.076	9.892	11.984
45	1.704	3.544	3.751	4.049	6.834	7.396	9.046	9.856	11.841
50	1.712	3.548	3.748	4.064	6.833	7.384	9.018	9.830	11.705
55	1.715	3.554		4.075	6.834		8.985		11.574
60	1.723	3.560		4.091	6.836		8.962		11.449
70	1.743	3.580		4.126	6.845		8.921		
80	1.766	3.609		4.164	6.859		8.885		
90	1.792	3.650		4.205	6.877		8.850		
95	1.806	3.674		4.227	6.886		8.833		
β mol/L	0.070	0.027	0.034	0.016	0.029	0.016	0.20	0.029	0.009

∗ A) Potassium tetraoxalate, 0.05 mol/kg; B) Potassium hydrogentartrate, satd. (25 °C); C) Potassium hydrogenphthalate, 0.05 mol/kg; D) Potassium dihydrogenphosphate, 0.025 mol/kg + disodium hydrogenphosphate, 0.025 mol/kg; E) Potassium dihydrogenphosphate, 0.009 mol/kg + disodium hydrogenphosphate, 0.03 mol/kg; F) Sodium tetraborate decahydrate, 0.01 mol/kg; G) Calcium hydroxide, satd. (25 °C); H) Potassium dihydrogencitrate, 0.05 mol/kg; I) Sodium hydrogencarbonate, 0.025 mol/kg + sodium carbonate, 0.025 mol/kg

Table 4. pH values of recommended technical buffer solutions (DIN 19267) [7]

	A	B	C	D	E	F
β, mol/L	0.26	0.11	0.10	0.09	0.07	0.33
0 °C	1.08	–	4.67	6.89	9.48	–
10	1.09	3.10	4.66	6.84	9.37	13.37
20	1.09	3.07	4.65	6.80	9.27	12.96
25	1.09	3.06	4.65	6.79	9.23	12.75
30	1.10	3.05	4.65	6.78	9.18	12.61
40	1.10	3.04	4.66	6.76	9.09	12.29
50	1.11	3.04	4.68	6.76	9.00	11.98
60	1.11	3.04	4.70	6.76	8.92	11.69
70	1.11	3.04	4.72	6.76	8.88	11.43
80	1.12	3.05	4.75	6.78	8.85	11.19
90	1.13	3.07	4.79	6.80	8.82	10.99

A) HCl, 0.1 mol/L; B) Glycin, 20 g/L + 10 g/L NaOH + HCl, 50 mmol/L; C) NaOH, 0.1 mol/L + acetic acid, 0.2 mol/L; D) $Na_2HPO_4 \cdot 12\,H_2O$, 55.4 g/L + citric acid monohydrate, 4.79 g/L; E) 1 L boric acid solution, 0.2 mol/L + 550 mL NaON, 0.2 mol/L; F) 0.1 mol/L NaOH.

Commercial secondary and technical buffers often contain chemical additives to prevent bacterial growth.

Operational pH standards. For preparation of *operational pH standards* (OS) [32] one needs only one reference value pH standard (RVS, one-point scale). This is solution C in Table 4.

An electromotive cell is employed that consists of two Pt | H electrodes and two liquid junctions, e.g.:

$$\mathrm{Pt} \mid \mathrm{H_2}, \mathrm{KCl}, c = 3.5\,\mathrm{mol/kg} \mid\mid \mathrm{RVS}, \mathrm{H_2} \mid \mathrm{Pt} \tag{2.8}$$

Since the hydrogen electrodes work with a theoretical slope, the pH is calculated according to Eq. (2.1):

$$\mathrm{pH} = \mathrm{pH_{RVS}} - \frac{U - U_{\mathrm{RVS}}}{k} \tag{2.9}$$

The *residual diffusion potential* (Section 3.3) is not eliminated. Table 5 lists some of recommended standards.

Nonaqueous Standard Solutions. All standard buffers are valid only for comparison in the same media. Aqueous solutions with extreme concentrations and nonaqueous solutions must be compared with buffers in a like solvent. For this primary and secondary standards with the same solvent as that of the sample are required.

Bates [33] used the Debye – Hückel equation to calculate the activity coefficients of chloride ions in alcohol – water mixtures. He introduced the constants A and B with the dielectric constant of the involved solvent. Since the ionic diameter $\mathring{A}$ in solvents is even less certain

than that in water, BATES also applied here the GUGGENHEIM convention (Section 1.3).

Table 5. Examples of operational pH standard solutions [32]

t, °C	A	B	C	D	E	F	G
0			4.664	4.729		8.399	9.475
5		3.466	4.657	4.722	6.477	8.238	9.409
10		3.470	4.652	4.717	6.419	8.083	9.347
15		3.476	4.647	4.714	6.364	7.933	9.288
20	1.475	3.484	4.645	4.712	6.310	7.788	9.233
25	1.479	3.492	4.644	4.313	6.259	7.648	9.182
30	1.483	3.502	4.643	4.715	6.209	7.513	9.134
37	1.490	3.519	4.647	4.722	6.143	7.332	9.074
40	1.493	3.527	4.650	4.726	6.116	7.257	9.051
50	1.503	3.558	4.663	4.743	6.030	7.018	8.983
60	1.513	3.595	4.684	4.768	5.952	6.794	8.932
70	1.52		4.713	4.800			8.898
80	1.53		4.75	4.839			8.88
90	1.53		4.80	4.88			8.84
95	1.53		4.83	4.91			8.89

A) Potassium tetraoxalate, 0.1 mol/kg; B) Sodium hydrogendiglycolate, 0.05 mol/kg; C) Acetic acid, 0.1 mol/L + sodium acetate, 0.1 mol/L; D) Acetic acid, 0.01 mol/L + sodium acetate, 0.01 mol/L; E) Piperazine phosphate, 0.02 mol/kg; F) Tris hydrochloride, 0.05 mol/kg + Tris, 0.01667 mol/kg; G) Disodium tetraborate, 0.05 mol/kg.

3. Electrode Cells

The following are used as pH indicator electrodes:

1) Hydrogen electrodes
2) Quinhydrone electrodes
3) Glass electrodes
4) Metal electrodes
5) Polymer menbane electrodes
6) pH-selective Field effect transistors (FET)

However, for industrial measurements today, the glass electrode cell is the instrument of choice.

3.1. Glass Electrodes

Historical Background. THOMSON [34] recognized in 1875 that glass is a solid electrolyte in which alkali metal ions can carry electrical current. This led CREMER [35] to propose that this substance might be an ideal semipermeable barrier for univalent cations, and in 1906 he used glass membranes in electrochemical determinations for the first time.

In 1909 HABER and KLEMENSIEWICZ [36] realized the relevance of the NERNST law and began to use the glass electrode for recording titration curves and other purposes [37]. They recognized the importance of a hydrated layer on the surface of the glass and took deliberate steps to produce it.

Construction. A glass electrode of typical design according to DIN 19263 [38] is illustrated in Figure 5. The spherical membrane at the lower end is the actual sensor. It is filled with *internal buffer* and fitted with a lead-off electrode, generally Ag/AgCl electrode. Since the glass membrane has a high electrical resistance, all the potential-bearing components are sensitive to electromagnetic fields and have to be shieled completely, inclusing cables and connectors. This is achieved by using a double-walled shaft containing metal foil.

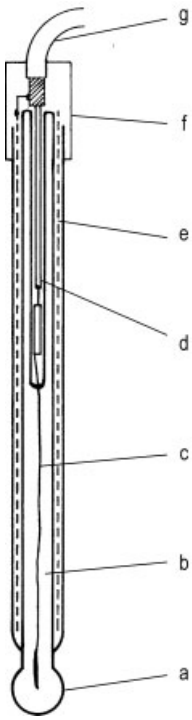

Figure 5. Standard glass electrode
a) Glass membrane; b) Internal buffer; c) Internal reference electrode; d) Lead connection space; e) Screening; f) Electrode head; g) Electrode lead

The Leached Layer. The surface layers of silicate glasses containing alkali metals undergo an ion-exchange process when they come into contact with water or aqueous solutions. Alkali metal ions from the glass go into solution and are replaced by hydrogen ions. A thin layer containing numerous hydroxyl groups is built up in this manner:

$$\equiv\text{Si–O–Na} + \text{H}_2\text{O} \rightarrow \equiv\text{Si–OH} + \text{Na}^+ + \text{OH}^-$$

The ion-exchange process commences as soon as a new or dry-stored glass electrode comes into contact with water. The exchange is followed by uptake of water to build up a hydrated layer. The hydroxyl groups allow the electrode to function as a sensor for hydrogen ions. However, ion exchange continues in the body of the glass, and so many alkali metal ions appear on the surface in the first few hours that it is only possible to obtain stable readings after ca. 12 h of immersion.

HUBBARD [39] found a conspicuous relationship between the hygroscopicity of glasses and their suitability for pH determination. An excerpt from his results is illustrated graphically in Figure 6.

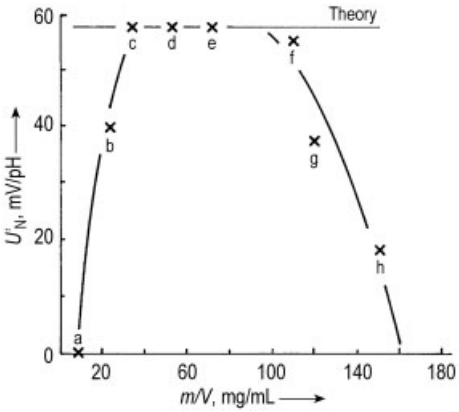

Figure 6. Relationship between water uptake of eight different soda-lime glasses (Na_2O–CaO–SiO_2 in wt %) and the slope of glass electrodes made of these glasses [39]
a) Quartz glass (0–0–100); b) Optical glass (15–25–60); c) Window glass (15–10–75); d) MacInnes glass (22–6–72); e) 25–5–70; f) 28–3–70; g) Waterglass (25–0–75); h) Waterglass (30–0–70)

BACH and BAUCKE [40] analyzed the surfaces of lithium glass membranes with a depth resolution of 3 – 5 nm by ion bombardement for spectrochemical analysis (IBSCA). The concentration profile in Figure 7 reveals the progress of the ion exchange during the first few hours and at various temperatures. The lithium concentration falls in a quite distinct manner as the surface is approached. Similar concentration profiles were discovered simultaneously by WIKBY [41] and DOBOS [42] on fractional etching with dilute hydrofluoric acid.

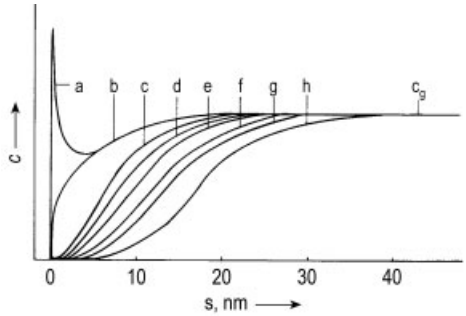

Figure 7. Concentration profiles of the outer layers of glass membranes with 25 % (LiO_2) after various conditioning times at 50 °C [40]
The left-hand side at $s = 0$ nm is the surface of the glass
a) Before penetration of water; b) Immediately after immersion and wiping off; c) After 10 min; d) After 30 min; e) After 1 h; f) After 3 h; g) After 8 h; h) After 16 h
16 h the leached layer reaches a steady state
c_g is the lithium concentration in the core of the glass

The degradation of glass does not end with the replacement of alkali metal ions by protons but continues from the surface inwards. With time, electrodes exposed to acidic or neutral solutions aquire a silic-acid-rich residue on their outer surface as a result of irreversible condensation.

$$\equiv\text{SiOH} + \text{HOSi}\equiv \rightarrow \equiv\text{Si–O–Si}\equiv + \text{H}_2\text{O}$$

This layer can be removed by drying, at pH < 3 by alkaline solutions at pH > 12, by hydrofluoric acid at pH < 3, and by mechanical rubbing.

The nonlinear concentration profiles in Figure 7 are the result of the differing diffusion rates of the ions in solution. It is known that in the case of the unchanged glass core, the alkali metal ions are relatively mobile, but protons are less so by a factor of 10^{-4} to 10^{-3} in the absence of incorporated water. BAUCKE and BACH [43] arrived at these orders of magnitude by means of experiments involving ionic migration in strong electric fields. SCHWABE and

DAHMS [44] achieved the same result using radioactively labeled tritium-containing solutions. In the leached layer the mobility of protons along hydrogen bond is much higher.

Membrane Glasses. The hydrogen-ion selectivity of the glass membrane depends on the ion exchange sites being compatible with protons with respect to charge, volume requirement, and coordination number. Membrane glasses should also be able to withstand much chemical and thermal conditions. The search for optimal glasses was complicated by the fact that each change in composition has a simultaneous effect on several properties. The membrane glasses in current use are always compromises for specific applications. There is no universally applicable glass of universal application.

MacInnes Glass (Corning 015) is a pure sodium calcium silicate glass (Na_2O 22, CaO 6, SiO_2 72 wt %) [45]. The first glass membrane suitable for a wide pH range was constructed from it. After its introduction in 1930 it became a standard glass which even today is still used for comparative purposes.

To extend measurements to higher pH levels and reduce the cross-selectivity towards sodium ions, the sodium ions in the membrane were replaced by lithium. The lithium ion is significantly more similar to the hydrogen ion than sodium ion with respect to ionic radius and coordination number.

The oxides of trivalent elements such as boron, aluminum and lanthanum are network formers which lead to glasses of high chemical stability.

According to PERLEY [46] lanthanum oxide can be added to membrane glasses to promote the lifetime at high pH values and temperatures around 100 °C without disadvantageous side effects on hydrogen-ion selectivity.

3.2. Reference Electrodes

The standard hydrogen electrode (Section 2.2) has a fundamental importance as a reference standard, but ready-to-use reference electrodes are preferred for practical determinations. A reference electrode according to DIN 19264 [37] is shown in Figure 8.

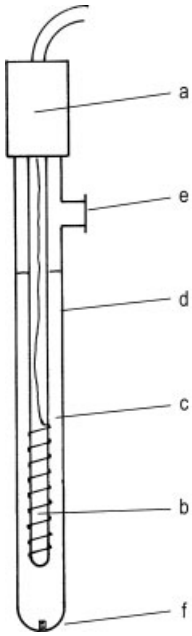

Figure 8. Standard reference electrode
a) Electrode head; b) Reference system; c) Reference electrolyte; d) Electrode shaft; e) Filling connection; f) Diaphragm

Table 6. Standard potentials of the most important Ag/AgCl reference electrodes [49], [50]

t, °C	*c* (KCl) at 25 °C			
	1 mol/L	3 mol/L	3.5 mol/L	satd.
0	249.3	224.3	222.1	220.5
10	244.4	217.4	215.2	211.5
20	239.6	210.5	207.6	201.9
25	236.3	207.0	203.7	197.0
30	233.4	203.4	199.6	191.9
35	230.4	199.8	195.4	186.7
40	227.3	196.1	191.2	181.4
50	220.8	188.4	182.4	170.7
60	213.9	180.3	173.5	159.8
70	206.9	172.1	164.5	148.8
80	199.9	163.1	155.6	137.8
90	192.7	153.3	146.8	126.9
95	189.1	148.1	142.5	121.5

The Ag/AgCl reference system is by far most frequently employed because it is simple to prepare and highly reproducible. Potassium chloride or more rarely cesium chloride is used as the electrolyte [48]. The potentials relative to the standard hydrogen electrode are accurately known; they are dependent on the concentration of the electrolyte and the temperature. The

most important reference electrodes with the Ag/AgCl system are summarized in Table 6.

Silver/silver chloride electrodes can be prepared electrolytically or thermally, depending on the properties desired. For temperatures above 80 °C, a tube is filled with a mixture of silver and silver chloride and stoppered with glass wool [51].

The classical *calomel reference electrode* [25] and the *Thalamid electrode* [52] are used more rarely, since their contents are highly toxic.

3.3. Liquid Junctions

In the boundary between two different electrolyte solutions, a potential arises which can be calculated by the *Henderson equation*, described in detail in [53].

Some potentials between the KCl solution of reference electrodes and sample solutions are listed in Table 7. It is not possible to obtain exact values since the structure of the diffusion zone is unknown. The diffusion potential causes the most uncertainty in pH measurements.

Table 7. Diffusion potentials U_d at 25 °C [10]

KCl //	HCl	E_d (mV)	KCl //	NaOH	E_d (mV)
satd.	1 mol/L	− 14.1	satd.	1 mol/L	+ 8.6
satd.	0.1	− 4.6	satd.	0.1	+ 5.4
satd.	0.01	− 3.9			
1 mol/L	1 mol/L	− 27.4	1 mol/L	1 mol/L	+ 18.8
1 mol/L	0.1	− 9.7	1 mol/L	0.1	+ 5.4
1 mol/L	0.01	− 2.8			

The junction between reference electrode and sample solution should always have as low a resistance as possible, but at the same time it should prevent mixing of the two solutions. These two contradictory requirements have led to more or less satisfactory compromises.

The best overview gives a free-flowing reference electrolyte, but in order to minimize the loss of electrolyte it is usually separated from the sample by a *diaphragm* (Fig. 9).

When a reference electrode is transferred from one sample or buffer solution to another, then the first sample that has already diffused into the junction still remains within it initially. The junction is then made up of the following strata: reference electrolyte | old solution | new solution.

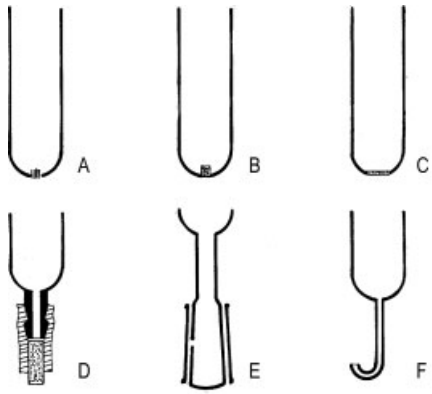

Figure 9. Designs of diaphragms
A) Asbestos or platinum wire; B) Porous ceramic; C) Extended porous ceramic; D) Exchangeable porous ceramic; E) Sleeve junction; F) Capillary

The sum of the diffusion potentials of a stratified junction is not equal to the diffusion potential of a direct junction. It takes several minutes for flowing electrolyte to entirely displace the old sample solution. So that for some time the cell exhibits a changing diffusion potential known as a *memory effect* [54]. A permanent outflow of 1 to 10 µL/d is recommended.

Eighty percent of the unexpected difficulties encountered with reference electrodes are attributable to the silver chloride dissolved in the electrolyte. It has been proposed to eliminate the complexed silver ions by an ion barrier in the exit of the tube containing the Ag/AgCl system. This can be a selective exchange resin or a metal which is less noble than silver [55].

The diffusion potential depends on the composition and the concentration of the reference electrolyte; 3.0 and 3,5 mol/L KCl are usually used. However, at temperatures below 0 °C additives are needed to lower the freezing point; an example of a *low-temperature electrolyte* is given in the following [6].

Water	250 mL
NaCl	65 g
Ethylene glycol	750 mL
$a(Cl^-)$	3.0 mol/L
fp	− 30 °C
bp	ca. 180 °C

The chloride ion activity must be verified by measurement with an electrolytic cell.

Diffusion potentials with extreme values do arise, at boundaries between two inmiscible solutions and when a reaction occurs between electrolyte and sample. Then a *double junction* is required with an electrolyte that is miscible with both solutions but does not react with them. That is known as an *electrolyte bridge*. Some examples of bridge electrolytes are listed in Table 8.

Table 8. Some bridge electrolytes recommended for nonaqueous solutions [56]

Solvent	Electrolyte	*c*, mol/L
Acetic acid, anhydrous	$LiClO_4$	0.1
Methanol	LiCl	1.0
Ethanol	LiCl	1.0
Organic solvent/water	$LiClO_4$	0.1

Reference electrodes with liquid electrolytes must always be kept under pressure so that sample solution does not penetrates into the electrode. This is not necessary when the electrolyte is made highly viscous or thixotopic by addition of thickening agents. Water-solouble agents which polymerize in the electrode are best [57]. However, it must taken in consideration that the sample can diffuse into the electrode without any impediment. Therefore, reference electrodes with *solidified electrolytes* have a limited lifetime.

Since all pH measurements are comparisons with standard solutions, the measured result always contains two diffusion potentials, which compensate each other to some extent. Considering this behavior, it would be accurate to write:

$$\Delta U_{\mathrm{d}} = U_{\mathrm{d}} - U_{\mathrm{dS}} \tag{3.1}$$

where ΔU_{d} is the *residual diffusion potential*. Its magnitude depends on the composition of the solutions involved.

The residual diffusion potential can be minimized by using calibration and sample solutions of similar composition. In the case of acids and bases, in particular, calibration and sample solutions should be of similar pH. For this reason, many buffer solutions have been developed for specific purposes.

A new surview of reference electrodes is given in [58].

3.4. Combined Electrodes

Glass and reference electrodes can be combined in a single rod. They are convenient and are used in most cases. Figure 10 illustrates a standard design of a combined electrode for laboratory use. The reference electrolyte is arranged around the glass electrode and serves as an electromagnetic screen. Ceramic and sleeve diaphragms (Fig. 9) are applicable in combined electrode.

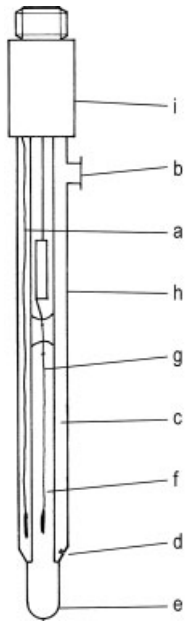

Figure 10. Laboratory combination electrode
a) Reference system; b) Filler opening; c) Reference electrolyte; d) Diaphragm; e) Glass membrane; f) Internal buffer; g) Internal electrode system; h) Electrode shaft; i) Electrode head

The concept of a combination electrode is only complete with the incorportion of a resistance thermometer, which is necessary for temperature compensation. The resistance thermometer is usually immersed in the reference electrolyte opposite the reference electrode.

Until the temperature equilibration of reference and internal glass electrode system has occurred, there will be particulary large potential deviations. According to [59], the deviation time can be minimized by constructing a thermally symmetrical electrode system and eliminating convection of the internal buffer. This is achieved by means of a common thermal isolation of the reference and internal electrode sys-

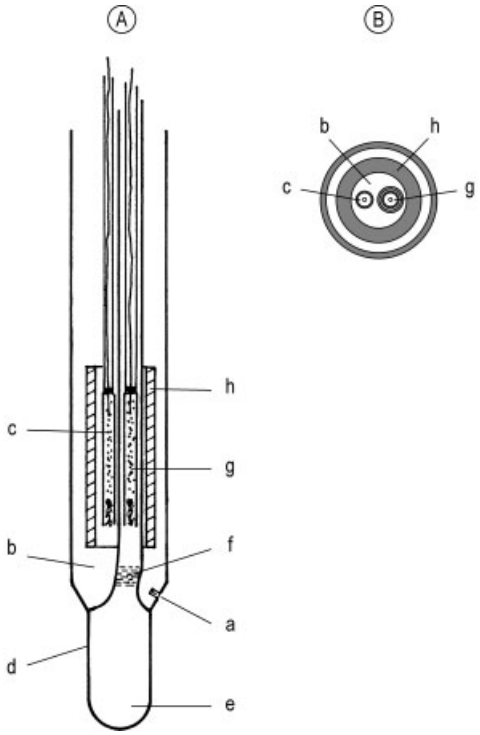

Figure 11. Lower part of a thermally symmetrical combination electrode
A) Longitudinal section; B) Cross section
a) Diaphragm; b) Reference electrolyte; c) Reference system; d) Glass membrane; e) Internal buffer; f) Convection brake; g) Internal electrode system; h) Insulating mantle

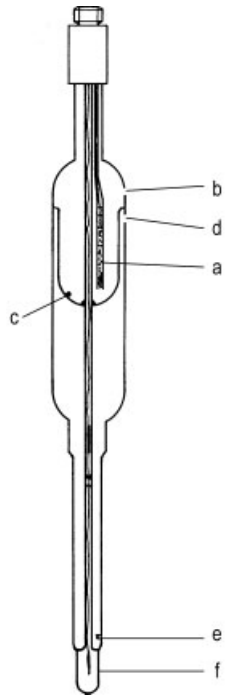

Figure 12. Combination electrode with electrolyte bridge for pressurized electrode assembly
a) Reference electrode; b) Filler opening for reference electrolyte; c) Internal diaphragm; d) Filler opening for bridge electrolyte; e) External diaphragm; f) Glass membrane

tems, as illustrated in Figure 11 (EQUITHAL system).

It is also possible to integrate an electrolyte bridge in the combined electrode (Fig. 12).

3.5. Testing Glass Electrode Cells

The reported *cell zero* pH_0 of a glass electrode only applies for a particular type of reference electrode and can only be tested in conjunction there with. It is the pH value at which the cell voltage is zero. Today's standard is a zero potential of the cell at pH = 7.0.

The correct determination according to IEC 746-2 [22] involves millivolt measurements in several secondary reference buffers which lie within the pH operating range of the glass electrode cell. The temperature should be maintained constant at $25 \pm 0.2\,°C$. The regression line, drawn or computed, gives the true pH_0'.

For cells with small deviation of *k'* a single buffer with a $pH(S_1)$ near pH_0 is enough. The simplified calculation of cell zero is:

$$pH_0\prime = pH\,(S_1) + \frac{U\,(S_1)}{k\prime} \tag{3.2}$$

According to DIN 19263 [38] the deviation of the measured cell zero pH_0' should not differ from the electrochemically defined pH_0 by more than $\Delta pH = \pm 0.5$.

The potential measured at the indicated cell zero is known as the *offset potential*:

$$U_{\text{off}} = k\prime\,(pH_0\prime - pH_0) \tag{3.3}$$

The *slope of a glass electrode cell* is the ratio of ΔU to ΔpH. In theory it should be equal to the NERNST potential *k* (see Section 2.2 and Table 2). The practical response *k'* is usually smaller. It is

obtained from the slope of the straight line Equation (3.2) can also be written

$$\mathrm{pH}_{0}\prime = \mathrm{pH}\,(\mathrm{S}_2) + \frac{U\,(\mathrm{S}_2)}{k\prime} \tag{3.4}$$

For routine measurements a second standard buffer solution S_2 with $pH(S_2)$ is used, such that $pH(S_1)$ and $pH(S_2)$ have a separation of two pH units. The slope is

$$k\prime = \frac{U\,(\mathrm{S}_1) - U\,(\mathrm{S}_2)}{\mathrm{pH}\,(\mathrm{S}_2) - \mathrm{pH}\,(\mathrm{S}_1)} \tag{3.5}$$

Temperature and pH range should be quoted in each case.

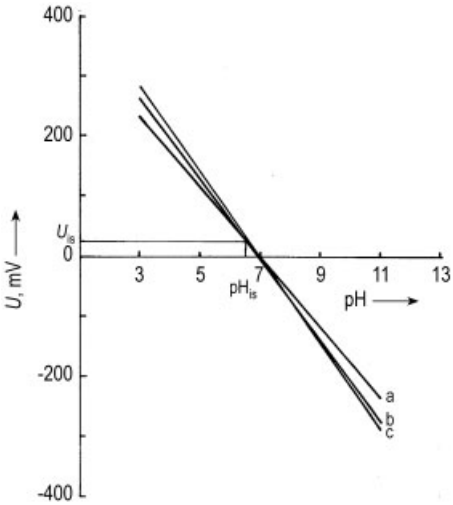

Figure 13. Graphical representation of the isotherm intersection
a) 25 °C; b) 65 °C; c) 85 °C

The *temperature coefficient* of the slope is defined as

$$\alpha = \frac{U_{\mathrm{T}_2} - U_{\mathrm{T}_1}}{U_{\mathrm{T}_1}\tau} \tag{3.6}$$

where U_{T_1} is the voltage of the cell at temperature T_1 and U_{T_2} at T_2. $\tau = T_2 - T_1$ is the difference between the two temperatures. The NERNST law gives the theoretical value, which at 25 °C is

$$\alpha\,(25...50^\circ\mathrm{C}) = \frac{1}{\tau}\left(\frac{T_2}{T_1} - 1\right) = 3.354 \times 10^{-3}\mathrm{K}^{-1} \tag{3.7}$$

The theoretical temperature coefficient is the basis of automatic temperature compensation of all pH meters.

The plotted slope following Equation (3.5) is a straight line with $\Delta U/\Delta\mathrm{pH}$ for a given temperature, it is therefore referred to as an *isotherm*. Isotherms at different temperatures have different slopes. The intersection of two or more isotherms is the *isopotential point* with the coordinates $\mathrm{pH}_{\mathrm{is}}$ and U_{is}. It is illustrated for an experimental example in Figure 13.

Using the four equations for the intersection with two standard solutions at two different temperatures gives equation 38.

Here indices 1 and 2 refer to one temperature, and 3 and 4 to the other; indices 1 and 3 refer to the first, and 2 and 4 to the second standard buffer solution. Since the straight lines intersect at a very small angle, the graphical determination of the intersection from two isotherms is less precise than its calculation. Knowledge of U_{is} is necessary for correct temperature compensation.

The *resistance of the glass membrane* at 25 °C rises with use at elevated temperature and with time. It gives information on the usefulness of the electrode. According to DIN 19263 [38] the maximal resistance should be $R_{\mathrm{n}} < 1000\,\mathrm{M\Omega}$ in the operating range.

The practical determination of the operating resistance of glass electrodes is carried out with direct current. However, the higher the potential applied, the more the glass membrane is polarized. The standard IEC 746-2 [22] therefore recommends using only the potential that is set up by the cell itself.

At first the glass electrode cell should be immersed in a solution which differs from the cell zero by $\Delta\mathrm{pH} \approx 2$. The potential is U_1. The second potential is measured after shunting a resistor with $R \approx 1000\,\mathrm{M\Omega}$ over the cell. The membrane resistance is calculated as follows:

$$R_{\mathrm{m}} = R_1\frac{U_1 - U_2}{U_2} \tag{3.9}$$

$$U_{\mathrm{is}} = \frac{(U_3\mathrm{pH}_4 - U_4\mathrm{pH}_3)\,(U_1 - U_2) - (U_1\mathrm{pH}_2 - U_2\mathrm{pH}_1)\,(U_3 - U_4)}{(U_1 - U_2)\,(\mathrm{pH}_4 - \mathrm{pH}_3) - (U_3 - U_4)\,(\mathrm{pH}_2 - \mathrm{pH}_1)} \tag{3.8}$$

According to IEC 746-2 [22] the *response time* required by the display to reach its final value after a sudden change in the sample solution can be divided into phases. These phases, illustrated in Figure 14, are 10 % time T_{10}, rise time T_r (or fall time T_f), and 90 % time T_{90}. It is not necessary to know what form the rate function takes to carry out this classification. To determine the percentage times it is only necessary to know the value of the final reading, the 100 % time.

The equilibrium time is primarily determined by the charge in the uppermost layer. Glass electrodes with thin leached layers react most rapidly. A conditioning time of one hour is optimal in this respect. As the electrode ages, the equilibrium time increases with increasing thickness of the leached layer and increasing membrane resistance. Soiled electrodes and those with old gel layers have response times up to 1000 times longer. Freshly etched electrodes have short equilibrium times. Equilibrium times mostly are shorter in the direction of positive potential change.

3.6. Alternative pH Electrodes

Glass electrodes are almost universal in their application but two alternatives for industrial applications should also be considered.

Enamel electrodes are the only example as yet of the successful reinforcement of a glasslike membrane. A portion of the insulating base enamel is coated with a thin layer of silver, which is then covered by the selective glass enamel [61].

Hydroxide-coated metals are electrodes for hydroxide ions. Since the hydroxide ion activity is related to that of the hydrogen ion by the ionic product of water (see Section 1.4), it is possible to use electrodes of this type to determine the pH. The sum equation can be written

$$Sb(OH)_3 + 3\,H^+ + 3\,e^- \rightarrow Sb + 3\,H_2O$$

However, the hydroxide concerned must not be readily soluble in acids or bases, and for this reason only the metalloids *antimony* and *bismuth* come into question for such electrodes [62].

The robust antimony electrode allows the continuous determination of pH under extreme operating conditions. Antimony electrodes can be continuously cleaned with brushes or corundum stones. The determination of pH under such conditions and in solutions containing hydrofluoric acid are the most important present-day applications of these electrodes.

Antimony and bismuth electrodes have other zero points and greater deviations of slope than glass electrodes. Therefore, they can not be combined with conventional pH meters. To calibrate cells with metalloid electrodes, secondary standard buffer solutions are too dilute. Technical buffers (Section 2.4) are better suited.

4. pH Meters or Transformers

4.1. Input Amplifier

Since the voltage drop resulting from connection of the input amplifier to the pH meter should not to be more than 0.1 % of the cell potential, the *input resistance* must be in the teraohm range, according to DIN 19265 [63] it must be greater than $5 \times 10^{11}\ \Omega$.

Accurate continuous pH determination can only be achieved over extended periods of time if the input voltage is modulated. pH meters transform the signal from the measurement cell into an alternating voltage before amplifying it and then rectify it again after amplification [64]. The alternating power supply feeds both the diode modulator and the synchronous rectifier.

The input amplifier is part of every pH meter. However, it may be advisible to take a second amplifier near the glass electrode as a *preamplifier* to avoid trouble with a long measuring cable.

4.2. Correction Functions

The most important function of pH meters is adjustment to the individual properties of the used electrodes.

The deviation of the cell *zero point*, ΔpH, is adjusted by a suitable compensation voltage. This voltage must be included in the temperature compensation (variable resistor R_2 in Fig. 15).

The *slope* k' in mV/pH is adjusted by varying the negative feedback of the amplifier (variable resistor R_3 in Fig. 15).

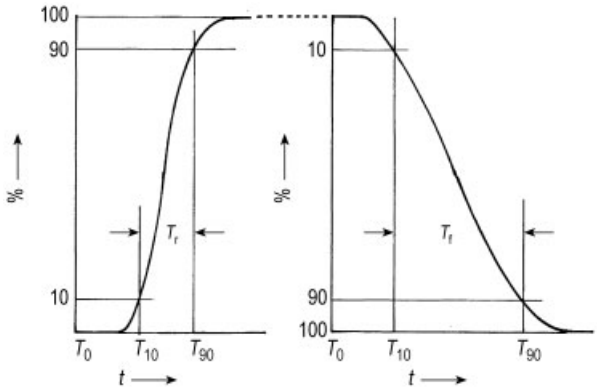

Figure 14. Rise curve [22]
T_0 starting time, T_{10} 10 % time, T_{90} 90 % time, T_r rise time, T_f fall time

The *temperature compensation* of a pH meter only takes account of the theoretical temperature coefficient of the NERNST theory (Section 3.5). The sample temperature is either adjusted by hand or automatically determined by means of a resistance thermometer. The resistance thermometer is a Pt-100 (DIN IEC 751) or a Pt-1000. The Pt-1000 is now preferred, since the cable resistance then can be neglected. The temperature compensation is based on the variation in the slope and is corrected with R_4 in Figure 15 in the same way. The output current flows through the temperature compensation resistor at up to 20 mA. To prevent a smaller current can be passed through the resistor and amplifiered in a separate transformation amplifier (not shown in Fig. 15).

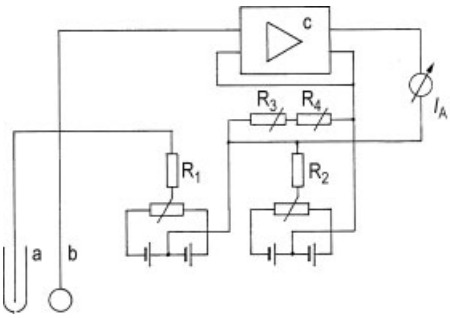

Figure 15. Circuit of an analogue pH meter [66]
a) Reference electrode; b) Glass electrode; c) Input amplifier; R_1) Variable resistor for U_{is}; R_2) Variable resistor for ΔpH; R_3 Variable resistor for mV/pH; R_4) Variable resistor for T

The compensation of the *isopotential point* by $\pm$ 100 mV should be possible (R_1 in Fig. 15).

4.3. Output

The output of the pH meter is a current of 0 to 20 mA or 4 to 20 mA at a minimum of 500 Ω. The output ranges are ΔpH = 5 and ΔpH = 10 [63]. In the case of 4 to 20 mA, the 4 mA signal be fed to a digital display.

4.4. Microprocessor-Controlled pH Meters

In digital pH meters equipped with microprocessors the digitized signal can then be processed according to established formulae. The cell signal is amplified as described in Section 4.1 and digitized by an analog/digital converter.

An advantages of these instruments is that none of the calibration buffers must have a value in the neighborhood of the cell zero, so that the calibration points can be chosen at will. The order of presentation of the buffer solutions can also be arbitary.

The digitalized raw data are automatically stored on calibration so that it is not necessary to adjust ΔpH and k'. The values of these variables are only calculated if they are called up for display. If pH values and voltages of two different standard buffers at two temperatures are stored, the isopotential U_{is} is calculated, too.

4.5. Testing pH Meters

To test a pH meter, it is necessary to provide an input that is independent of the conditions of the electrodes and their leads; a so called *pH simulator* is used for this purpose. It produces the same voltage as a pH measurement cell but by purely electronic means. It consists basically of a constant voltage source and a potential divider [22].

When the input is short-circuited the *offset voltage* of a pH meter is directly displayed in the mV range as output U_{A0}. Another output voltage U_{A1} is obtained if the short circuit is replaced by a high-value resistor ($R \approx 1000\,\mathrm{M\Omega}$). The offset or *interference current* is given by:

$$I_{\mathrm{off}} = \frac{U_{A1} - U_{A0}}{R} \tag{4.1}$$

It may amount to only a few picoamperes.

4.6. Self Monitoring

Despite all the progress in pH measurement, glass electrodes with their thin membranes and reference electrodes with open diaphragms are sensitive elements. There are some methods to supervise the electrodes by electronic means. The function can monitored by automatic comparison of the actual pH value with two stored values of the calibration buffer solutions. Unexpected large changes produce an alarm [68].

Membrane break ages can be detected by means of the resulting drop in diminition resistance.

Blocking of the diaphragm and the associated increase in the resistance of the reference electrode is one of the most frequent problems. This is monitored by impedance measurement with alternating current. Interferences result from the capacitance of the electrodes and cables. This can be eliminated with a test circuit containing a capacitance in series. The phase of the test current is analyted by variation of the frequency [70]. It is performed without interrupting the measurement.

Variations in the impedance of the sample solution can disturb the impedance measurement. The electrode impedance is compensated by simultaneous measurements of the sample impedance [72].

4.7. Self Calibration

The pH values of the most important *standard buffers* in the usual temperature range can be stored in a microprocessor-controlled pH meter so that they can be selected afterwards without the risk of operator error. Some manufacturers have gone even further. The instrument itself searches for the used buffer by assuming maximum cell-zero deviations of, say, $\Delta\mathrm{pH} < 0.5$. The instrument monitors the temperature and searches for the correct pH of the buffer solution for the particular temperature [72]. In the case of two-point calibrations the buffer used can also be identified by means of a plausibility test of the calculated electrode data [73].

It is only a small step for a microprocessor controlled pH meter to *self-calibration*. The problem lies primarely with the electrode assembly and the control apparatus (see Section 6.5). The pH meter must be able to carry out the adjustment of $\mathrm{pH}_{0'}$ and k' automatically on receipt of a signal from the control apparatus.

Membrane breakage and diaphragm blockage happen rarely. More frequently, zero point and slope of a glass electrode cause trouble. MELZER [74] proposes a circuit with two glass electrodes, one reference electrode, and two pH meters. The two glass electrodes contain different internal buffers, e.g., with $\mathrm{pH}_{01} = 4.6$ and $\mathrm{pH}_{02} = 7.0$. Then the output with one sample solution with pH is:

$$\begin{aligned} U_1 &= k\prime\,(\mathrm{pH}_{01} - \mathrm{pH}) \\ U_2 &= k\prime\,(\mathrm{pH}_{02} - \mathrm{pH}) \\ U_1 - U_2 &= \Delta U = k\prime\,(\mathrm{pH}_{01} - \mathrm{pH}_{02}) \end{aligned} \tag{4.2}$$

This means that the voltage difference is independent of the sample pH, but it is proportional to the slope, supposed the zero points pH_{01} and pH_{02} remain constant and the slope of both the electrodes are equal. The slope can be measured and compensated on-line if the pH meters are designed properly. The sensor need not be removed from the measured liquid, and a standard buffer is not necessary. Unfortunately on-line determination of zero points is not possible with this method.

Membranes and lead off electrodes should be of the same type. This is best achieved with a

double glass electrode. For the design of a combined electrode for this method, see Figure 16.

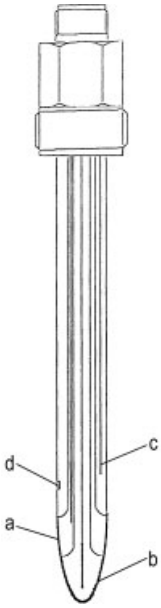

Figure 16. Combined double membrane glass electrode
a, b) Glass membranes; c) Reference electrode; d) Diaphragm

5. Analytical Applications

5.1. Calibration of Glass Electrode Cells

Two-Point Calibration. In contrast to Equation (3.5) it is sufficient to use two buffer solutions, at least for routine calibration. Both measurements that make up the calibration must be made at the same temperature, ideally at the temperature at which the samples will be investigated. If there is a manually adjusted temperature compensation this should initially be set to the temperature of calibration. The temperature should be maintained within at most ± 0.5 K of the calibration temperature.

The zero point is adjusted by using the first buffer that is as close as possible to the reported cell zero. This step is expressed in Figure 17 in terms of a parallel displacement of the calibration line along the abscisa by $\Delta pH = pH_0' - pH_0$. The pH of the second buffer solution should differ from pH_0 by at least 2 pH. It is used to adjust the slope of the straight line k' (in mV/pH). The two calibration solutions should bracket the desired measurement range. The adjustment of the slope also affects the adjustment of the zero point. If the pH_1 of the first buffer differs from pH_0' by more than 0.5, then the calibration must be repeated.

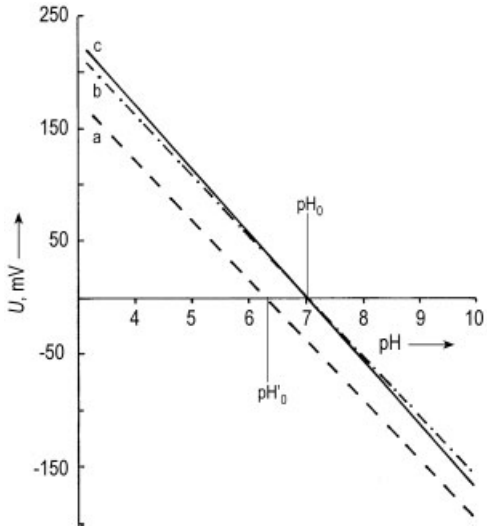

Figure 17. Correction of zero point and slope
a) True relationship between pH and cell voltage; b) Relationship in pH meter after adjustment of ΔpH; c) Variation after adjustment of k' in mV/pH
pH_0 quoted cell zero, $pH_{0'}$ actual cell zero

The standard deviation of the two-point calibration amounts to $\Delta pH \approx 0.02$.

Multipoint Calibration. If more than two calibration points are plotted they will not lie exactly on a straight line. DIN 19268 [76] and Baucke et al. [77] therefore recommend *multipoint calibration* with three or four calibration solutions in cases where high accuracy is required. For this the millivolt range of the pH meter is used. Each reading should be made when the drift has settled to less than 0.1 mV/10 min. This can mean that waiting times of up to 30 min can be required for each determination. The most probable straight line is than calculated from the results by the method of least squares [78], [79].

Uncertainity of pH Measurement. In this section ΔpH values of 0.01 to 0.02 were quoted, depending on the calibration method. This is valid only under pure laboratory conditions. For industrial environments McMillan [80] esti-

mated uncertainities of $\Delta pH \approx \pm 0.25$. In case of multipoint calibration it is possible to calculate the mean error [78].

5.2. pH Measurement in Water Analysis

The recommendations of DIN 38404-5 [81] and ISO 10523 [82] are applicable to drinking, process, spring, and surface water with conductivities of $\sigma > 100\,\mu S/cm$. Conventional sampling techniques and standard buffer calibrations can be used. According to the *German drinking water law* (Trinkwasserverordnung) [83] the pH of ready-to-use drinking water must lie between 6.5 and 9.5. If the public mains contain asbestos cement then in addition the pH must not deviate by more than $\Delta pH = 0.2$ below the pH of *calcium carbonate saturation* [84].

Difficulties are often met in measuring the pH of samples with a low conductivity. This is particulary true of high-purity water, produced by distillation, deionization, or dialysis, and rainwater. For the last-mentioned see [85].

Galster [86] has made some effective recommendations for completely desalted water on the basis of theoretical considerations and practical experience: It is preferable to use a reference electrolyte containing not more than $c(KCl) = 1$ mol/L and a sleeve as-liquid junction (Section 3.3). Since electrostatic charges take some time to flow away it is recommended to provide the sample solution with a shield and a large-surface ground.

The extremely small buffer value means that the pH of a sample drops rapidly as carbon dioxide is absorbed from the air.

The response of glass and reference electrodes have a hysteresis effect when they are changed from usual buffer solutions into a low-ionic water. It is better to employ diluted standard buffers [87]. Technical buffers should not be used.

In wastewater the danger of contamination of the electrodes is greater and the accuracy that is required is lower. For this reason it is best to use reference electrodes with extended or exchageable diaphragms (Section 3.3). Reference electrolytes with polymer-immobilized reference electrodes and an open junction are permitted [88].

5.3. pH Measurement in Concentrated Solutions

Nonelectrolytes. It has long been known that in concentrated solutions of sugars low measured pH values are too high and high measured pH values are too low when compared with calculated values. A solution containing 26 wt % sucrose exhibits an apparent slope of only 53.2 mV/pH between pH 3 and 12 and the diffusion potential is elevated [89], [90]. This effect is the result of changes in the fine structure of water caused by the presence of the sugar molecules. Highly concentrated solutions can no longer be regarded as purely aqueous, so that it is necessary to take into account medium effects (Section 1.4).

Electrolytes. Schwabe [91] has made extensive investigations of pH displacement in concentrated solutions of electrolytes. The lowering of pH of a acid solution by the addition of electrolytes follows by the enlargement of ionic strength. The diffusion potential in concentrated solutions is lower than calculated as in Section 3.3, because the rates of migration of ions are diminished at concentrations $b > 1$ mol/kg. Conventional glass electrodes are suitable for such determinations. Hydrogen and glass electrodes agree closely with each other even at very high concentrations and at temperatures above 100 °C [92].

Strong Acids. The major problem of pH determination in strong acids is the magnitude of the diffusion potential. The diffusion potential can be stabilized by interposing an electrolyte bridge between reference electrode and sample containing sample solution (see Section 3.3).

The cross-sensitivity of glass electrodes towards anions was mentioned with in Section 3.5, the "acid error" only appears after several hours; therefore, it is recommended that individual measurement be made rapidly.

Strong Bases. Potassium hydroxide solution allows the production of solutions with $pH > 14$. However, in highly concentrated solutions a much of the water is required for hydration, so that the *activity of water* can drop. The activity

coefficient of hydroxide ions $\gamma(OH^-)$ increases to much as 30 in concentrated hydroxide solutions. As the hydroxide ion concentration increases, the pH increases in the following way:

$$\mathrm{pH} = \mathrm{p}k_a - \log a\,(\mathrm{H_2}O) + \log b\,(\mathrm{OH^-}) + \log\gamma\,(\mathrm{OH^-}) \quad (5.1)$$

This means that a concentrated solution of, e.g., potassium hydroxide having a concentration of 17.5 mol/kg has a theoretical pH of 17.65 [93].

However the cross-sensivity of the glass electrode towards cations is a much greater problem (Section 3.5).

Even the best glass electrode must be expected to deviate greatly above pH 14.

5.4. Suspension Effects

Potentiometric determinations in suspensions, sludges, and particulary in soils can exhibit unexpectedly large errors. This effect is also known as the *Wiegner – Pallmann effect* [94], [95]. Two independent phenomena are involved [96].

1) The *type 1 suspension effect* influencing the sensor electrode
2) The *type 2 suspension effect* influencing the reference electrode on account of an anomalous diffusion potential.

The type 1 suspension effect was initially demonstrated in investigations on suspensions of γ-Al_2O_3 in pure water (Fig. 18). The apparent pH measured in the suspension was about 0.4 pH units higher than in the clear supernatant. Aluminum oxide and some other substances absorb hydrogen ions from acid solutions and hydroxide ions from basic solutions and thus act as a suspended buffers.

The type 2 suspension effect occurs more frequently than type 1 and its magnitude is up to ten times higher. It is especially to be observed when substances possess ion-exchange properties. Figure 18II demonstrates the effect schematically.

The relatively immobile suspended particles, which are electrically charged on account of the zeta potential, reduce the mobility of ions with like charge and hence have an influence on the magnitude of the diffusion potential as soon as they come into contact with reference electrolyte [98]. The effect also must be considered in suspensions of microbiological fermenters [99].

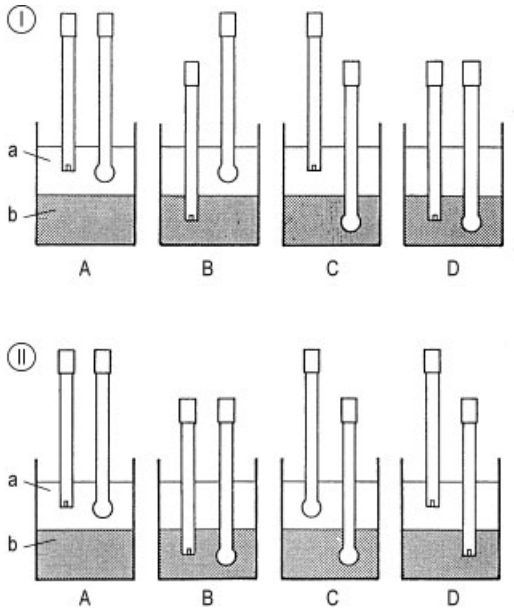

Figure 18. Suspension effects
I) Type 1 suspension effect: the diagrams indicate the measured pH as a function of the position of the reference and glass electrodes in the clear supernatant and in the sediment
A) pH = 8.56; B) pH = 9.57; C) pH = 8.91; D) pH = 8.92
II) Type 2 suspension effect: the diagrams indicate the measured pH as a function of the position of the reference and glass electrodes in the clear supernatant and in the sediment
A) pH = 6.0; B) pH = 2.0; C) U = 0 mV; D) U = 240 mV
a) Suspension solution; b) Suspension

The type 2 suspension effect is dependent on the composition of the reference electrolyte. It can reduced by including an electrolyte bridge with an electrolyte consisting of a clear solution obtained by filtration or centrifugation of the sample.

5.5. pH Measurement in Nonaqueous Solvents

Liquid ammonia is the best investigated nonaqueous inorganic solvent. The use of liquid ammonia as a solvent for the titration of very weak acids has been suggested [100]. The glass electrode can be used for determinations below the boiling point of liquid ammonia (– 33.38 °C). For this purpose a high-conductivity membrane

glass is combined with a low-temperature internal buffer. The authors used a cell in which liquid ammonia is also the solvent for the reference electrolyte. The used glass electrode has a mixed solvent with methanol as internal buffer and is conditioned as usual in aqueous solution and immersed in liquid ammonia after it has been dried briefly.

Cruse [101] has provided a comprehensive description of the determination of pH in organic solvents. Figure 1 shows the approximative relative positions of the pH scales in various solvents [17]. However, it is not possible to make direct comparisons of pH values on different scales. The correct procedure would be to set up a pH scale with primary pH standards for each solvent, as described in Section 2.4. Nonaqueous standards have been published for alcohol – water mixtures [102]. For the preparation of standard buffers in other solvents, see [103].

The glass electrode is useful in nonaqueous solutions, provided it possesses an leached layer which was grown in aqueous solution. It is necessary to regenerate it after some time. Conventional reference electrodes with aqueous potassium chloride can be employed if the solvent is miscible with water. The residual liquid junction potentials are not greater than in aqueous solutions. In some cases, specifically when the sample is immiscible with the reference electrolyte, it is recommended to use an electrolyte bridge with an organic electrolyte solution (Section 3.3).

Here general aspects of pH measurements are described. Numerous examples for laboratory and field measurements be found in [104].

6. Technical pH Measurements

6.1. Extreme Conditions

Diffusion potentials and cross-sensitivies become smaller at *low temperatures*. It is the increase in the resistance of the membrane that causes some problems; this resistance approximately doubles for each 10 K that the cell is cooled. The lower limit of commercial electrodes is − 30 °C. The internal buffer of the glass electrode must contain solvents that act as antifreeze.

The reference electrode requires a reference electrolyte with a low freezing point [105].

The temperature compensation range in most pH meters ends at 0 °C. Sometimes it is possible to make a correction by adjusting k' by means of the slope-correction potentiometer.

The *upper temperature limit* for present-day glass electrodes is 150 °C and lies in the nature of glass on account of the onset of decomposition in aqueous solutions and recrystallization.

Above 150 °C zirconium oxide and palladium hydride electrodes are suitable [106]. With yttrium-stabilized zirconium oxide electrodes measurements are possible in supercritical aqueous systems [107]. For calibrations at temperatures above 100 °C, pH values of standard buffers up to 200 °C are available [62].

Standard glass electrodes withstand a *pressure* of up to 9 MPa. Reinforced glass electrodes can withstand external pressures of up to 15 MPa. Below this pressure, the only complication is the pressure-compensation system for the reference electrode.

If it is only necessary to equilize a few meters of water gauge, then it is sufficient to raise the electrolyte reservoir to an elevated position. For higher pressures, a *pressure chamber assembly*, such as that illustrated in Figure 19 is used. A cylinder, 19 or 25 mm in diameter, equipped with an O-ring for immersion into a matching support, has become standard. A constant inner pressure is limited by the ridity of the reference electrode to 0.3 MPa according DIN 19264 [47]. Higher pressures are possible by pressure control (see Section 6.4).

Pressurized assemblies are equipped with combination electrodes to save space; the electrolyte reserve is accommodated in a cylindrical anlargement of the electrode stem.

In biotechnology sterilization of pH electrodes is necessary. Heat sterilization with saturated steam at 130 °C and up to 0.3 MPa is unavoidable in microbiological applications [108]. The projecting portion of the reference electrode containing the major portion of the electrolyte is not sterilized. It must therefore be assumed that the electrolyte does not contain harmful bacteria, or that these cannot penetrate the diaphragm into the sample solution. Only ceramic diaphragms with mean pore sizes of less than 1 µm are permitted. In addition, the seal between the pressure

chamber and the electrode must not allow leakage of air.

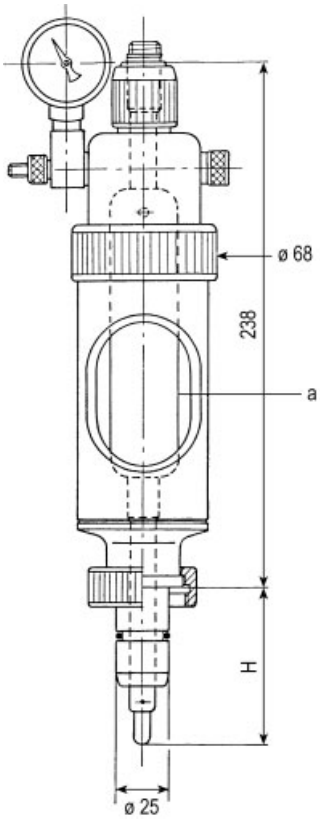

Figure 19. Pressure-chamber electrode assembly (Mettler-Toledo)
a) Combination electrode

6.2. Installation

The high input resistance of a modern pH meter ($R_e > 1\,T\Omega$) is of no avail if the electrical connections of the lead to the glass electrode do not exhibit insulation leakages on this order of magnitude. The insulation resistance of a polyethylene-insulated glass electrode lead is about $100 \times 10^{15}\,\Omega$/m, so that leakage is not expected, even over long runs. However, since the capacitance is 50 – 100 pF/m the cable should be kept as short as possible and installed without avoidable convolutions. All connections must be absolutely watertight.

The glass electrode lead must be shielded completely. In combined electrodes it is permissible to connect the shield of the cable to the reference electrode without earthing. For industrial applications, the shield cable should be earthed following the connection. Since strong induction currents are possible, the lead should be installed in a earthed steel conduit when the runs are long. Since shielded cables are sensitive to movement, they should not be installed on vibrating supports.

Electrical fields can also be set up within the sample solution if stray electrical current flows through it. This source of error is avoided by placing the electrodes within a Faraday cage.

In industrial plants the sample solution is generally earthed. Determinations in low-conductivity solutions and with glass or plastic electrode supports require their own earth. This consists of a massive rod of stainless steel, connected to the casing of the pH meter. To avoid a ground loop, the lead must be installed jointly with the electrode cable. For installation, see [109].

6.3. Electrode Assemblies

Immersion Assemblies. Immersion electrode assemblies, such as that illustrated schematically in Figure 20, are used for determining pH in open vessels. Since the electrodes are immersed below the surface of the liquid the hydrostatic pressure must be compensated by raising the level of the electrolyte vessel. There is no need for a reservoir if the assembly is equipped with a sealed reference electrode containing solidified electrolyte.

The most important application of immersion assemblies is in wastewater treatment facilities where they are mounted in open channels and basins. GÜGEL [110]. A detailed study of the immersion assemblies of several manufactures has revealed that cleaning and maintenance need only be carried out every four weeks and that a measurement uncertainty of $\Delta pH = 0.15$ is to be expected [111].

Built-In Assemblies. Built-in electrode assemblies are used in closed containers. The simplest consist of a mounting plate, like the lower end of an immersion unit, or a simple screw adapter accommodating a combination

electrode. In most cases, pressure compensation is necessary.

If the unit is installed in a vertical wall the electrodes should have an angle of at least 15° with respect to the horizontal, so that the reference electrolyte and the internal buffer will collect in the lower portion of the electrodes.

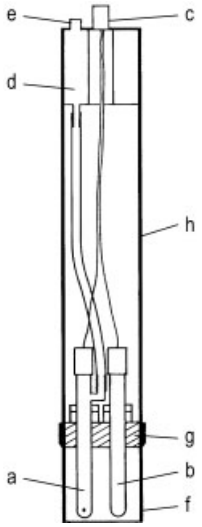

Figure 20. Schematic diagram of an immersion electrode assembly
a) Reference electrode; b) Glass electrode; c) Socket connection; d) Electrolyte reservoir; e) Filler; f) Protective cage; g) Mounting; h) Shaft

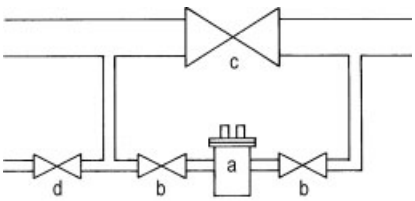

Figure 21. Bypass for a flow-through cell
a) Flow-through cell; b) Valves for maintenance; c) Throttle valve; d) Valve for taking samples

Flow-Through Assemblies. Flow-through electrode assemblies are built-in assemblies that are mounted in small vessels and installed in pipelines. The sample stream flows past them. If possible, the sample should flow freely after it has passed the electrode unit so that pressure compensation is unnecessary. Measurements in closed systems require the use of a *bypass* as illustrated in Figure 21. This electrode unit can be removed for maintenance after the valves have been closed.

6.4. Pressure Control

A pneumatic pressure control circuit for an electrode unit is illustrated in Figure 22. A tube for the removal of condensate is also required if the sample solution has a high vapor pressure.

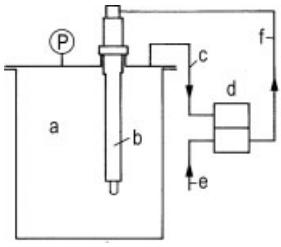

Figure 22. Pressure control of an electrode assembly
a) Pressure vessel; b) Pressurized electrode assembly; c) Control pressure; d) Differential pressure relay; e) Compressed gas supply; f) Top-up pressure supply; P) pressure gauge

A simple arrangement of Figure 23 can be used instead of a differential pressure relay [112]. This device is also applicable in cases where the vapor pressure is high or where aggressive gases are encountered. A small constanty controlled stream of air is blown into the sample vessel through a throttle.

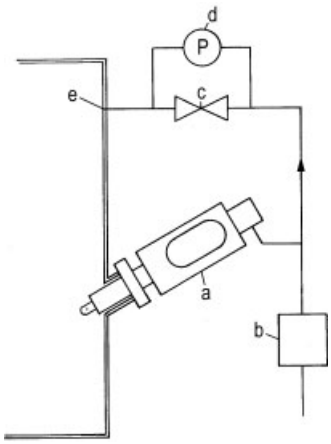

Figure 23. Device for ensuring constant excess pressure in a pressurized electrode assembly [112]
a) Pressurized electrode assembly; b) Flow regulator; c) Choke valve; d) Pressure difference display; e) Air entry line

6.5. Retractable Assemblies

When the electrode assembly is installed in a situation where it can not be dismounted without shutting down the plant, it may be advisable to install a retractable electrode assembly. The constructions in general use employ combination electrodes with a pressure chamber.

Figure 24A illustrates an insert with a sliding cylinder. When slid back, as in Figure 24B the closed cylinder form a seal in the bore in front of the electrode assembly. The electrode assembly can now be removed while cylinder seal remains in its seal. In this position the space between electrode and cylinder can be filled with different solutions. It is possible to clean the electrode with chemicals and to recalibrate with buffer solutions. When the electrode has been tested it can be returned to the measuring position.

Instead of closing the cylinder with an O-ring, one can supplement the assembly with a valve [113].

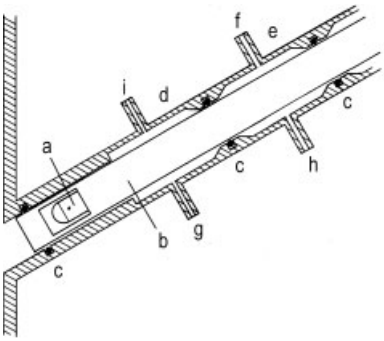

Figure 25. Interchangeable electrode head with divided outer chamber in inspection position [116]
a) Combination electrode; b) Insert; c) Cylindrical seal with O-rings; d, e) Outer chambers; f), g) Outer chamber outlets; h, i) Outer chamber inlets

Retractable electrode assemblies can also be operated pneumatically. The design illustrated in Figure 24 is particularly suitable for this purpose. Operational errors can then be avoided by suitable design of the pneumatic system.

Bioreactors are usually equipped with pressurized-chamber electrode units. In the course of the process it may be necessary to clean and recalibrate the electrode. Steam treatment of assembly in the retracted position does not guarantee sterility. Whatever the construction and sterilization conditions, some parts of the electrode assembly remain inaccessible to the sterilizing action of the steam, and the chemicals and pressurizing air can transport microorganisms into the nutrient. SCHAAD et al. [116] suggested solving this problem with a divided chamber, as illustrated in Figure 25. Both portions can be sterilized with saturated steam. When the electrode is thrust down after sterilization, the front part of the divided chamber is only exposed to already sterilized portions of the electrode assembly. A microbiological filter in the connection to the air line then makes it possible to top up the pressurize.

7. pH Control

General Remarks. It is necessary to keep the pH constant during many industrial processes. Irreparable damage can be caused by brief deviations of the pH, particulary in the case of microbiological processes.

The pH is controlled by comparing the actual value x(pH) of the controlled variable with the desired value w(pH). The desired value can be maintained constant as in a continuous process or it can be altered according to a time table as in a batch process. Control is exerted by a closed circuit known as control loop. This is represented graphically by means of a block circuit diagram, such as that illustrated in Figure 26. Four different functional units can be distinguished in this figure. A control loop is in principle unstable and can oscillate. The greater the influence of the controller on the control parameter, the less stable is the control.

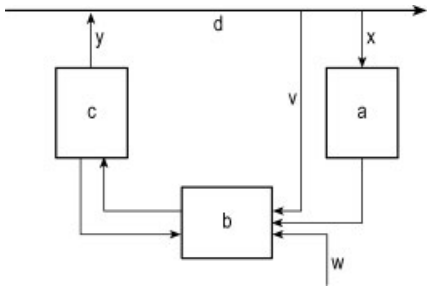

Figure 26. Block diagram of a control loop
a) pH meter; b) Controller; c) Control element (valve, pump); d) Controlled section
x(pH) controlled variable, w(pH) reference value, y controller output, v feedback

For informations on general control, see → Process Control Engineering, Chap. 4.

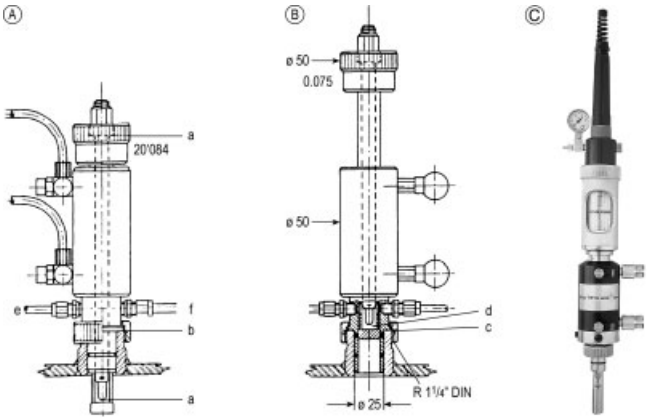

Figure 24. Detail of a retractable electrode assembly with cylindrical seal (Mettler-Toledo)
A) Measuring position; B) Withdrawn position
a) Combination electrode with solidified electrolyte; b) Insert; c) Cylindrical seal with O-ring; d) Anterior chamber; e) Anterior chamber inlet; f) Anterior chamber outlet
C) Photoshowing with the pressure chamber (Mettler-Toledo)

pH Functions. The *stationary gain* Ks is the dependence of the

$$K_s = \frac{\Delta x\,(\mathrm{pH})}{\Delta y\,(q)} \tag{7.1}$$

reagent input x(pH) on the controller output $y(q)$ [117]. In pH control Ks is never a constant value. Its plot as a function of pH is known as a *titration curve*. The slope of pH curves can vary by factors of up to 10^4. Accurate prior knowledge of the titration curve is a prerequisite for pH control.

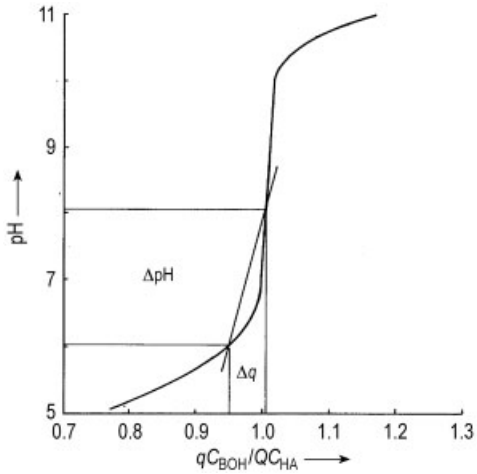

Figure 27. Titration curve with approximate straight line
Sample 100 mL, acetic acid $c = 1$ mol/L, titrated with c(NaOH) = 1 mol/L

That part of the titration curve which can be expected to encompass the range of the control variable (*control range*) is approximated by a straight line, as illustrated in Figure 27. Although this simplification is sometimes very inaccurate, there is unfortunately no way in which it can be avoided when planning a pH control system. The reciprocal gradient of this straight line is equal to the buffer capacity for flowing systems (cf. Section 1.2)

$$\beta_Q = \frac{\Delta q}{Q \Delta \mathrm{pH}} \tag{7.2}$$

Q is the main stream in L/min and q the addition in mol/min.

For controllability, the response time within the reaction vessel is important. The *step response* gives the progress of pH without any control after the addition of a strong acid or base in one step (Fig. 28). Its form depends on the chemistry and geometry of the reaction vessel. One can distinguish time sections between dead time T_d, *delay time* T_del, and *capacity* or time constant T_c. The rate

$$I_\mathrm{v} = \frac{T_\mathrm{c}}{T_\mathrm{d} + T_\mathrm{del}} \tag{7.3}$$

is decisive for the controllability of the process.

Discontinous Control. *Two-point control* has only two positions for the output: "on" and

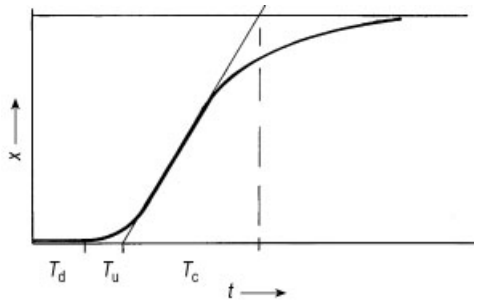

Figure 28. Response function. T_d dead time, T_u delay time, T_c time constant

Figure 29. pH transmitter 2500, (KNICK, Berlin).

"off". Such control of pH can be simply and cheaply achieved by installing a magnet valve.

An on – off controller can never maintain the pH at a truly constant value; instead it sets up periodic oscillations. For a comprehensive theory of an on – off control, see [119]. Many pH control requirements can be fulfilled by means of an on – off controller together with a suitable timer the determines the closed and the open times. KUBLI [119] achieved with optimal adaption of t_{on} and t_{off} a ΔpH of ± 0.02 in microbiological reactions.

If it is necessary to control a chemical process by the addition of acid or base, depending on the pH (e.g., when neutralizing wastewaters), it must be possible to activate two control devices at will. Two independent on – off controllers or one three-point controller can be employed for this purpose.

The three-point controller is normally employed in conjunction with a continuously variable control device, e.g., an electric motorized valve or a variable pump. This is a *quasicontinous control* system which has an integral characteristic, since the valve sums all impulses it receives. However such a system is unstable in pH control. The pure integral characteristic of three-point control can be converted by means of feedback. Figure 29 illustrates an industrial pH-meter with incorporated automatic calibration and a three-point controller.

Feedback. It is possible to diminish the amplitude of on – off control by means of a feedback mechanism. This takes the form of an electrical model of the controlled volume made up of capacitors and resistances in abbreviated timescale. The model is wired parallel to and in the opposite sense to the controller. The feedback is applied with a delay time and with a reduced time. After a short period the feedback is slowly reduced and switched off until the controller is actuated once again. This sequence is repeated until the pH deviation disappears.

The control variable responds rapidly at first, and then when it exceeds the level of the proportional adjustment it is only switched on in pulses. This means that the valve setting approaches its new value gradually. The properties of this form of control are very similar to those of a proportional integral differential (PID) controller (see Fig. 30).

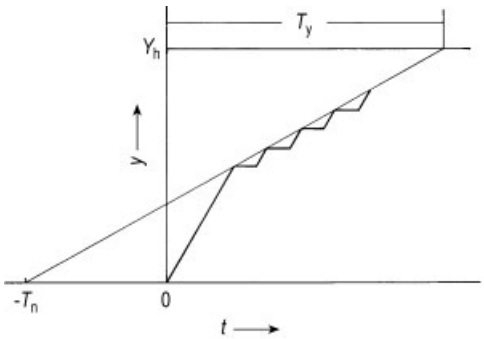

Figure 30. Variation of the output of a three-point controller with delayed reducing feedback

The three-term controller with feedback is the most important device used in pH control. It is in no way inferior to a continuous controller. It also delivers the pulsed output appropriate for electric motorized control valves. Since the control

steps are dependent on the impulse width, such controllers are sometimes referred to as *impulse width controllers*.

Feedforward Control. In a simple control circuit, errors of long duration cannot be avoided if, e.g., the input of the supply stream exhibits sudden large changes in flow rate. This is because the pH control system must respond slowly to the pH change for reasons of stability. A *feedforward system* is resorted to in such cases. The rate of flow of the input stream, or other important influence, is monitored and the signal obtained is used to bring about an immediate and proportional change in the rate of addition so that large deviations in the pH are prevented from occurring [120].

Direct feedforward causes the controller output to be multiplied by the relative change in the disturbing variable. In the case of on – off and three-point controllers this can be realized by adjusting the on-time.

Feedforward control can be carried out very elegantly when the pH is controlled by means of a dosing pump, whose stroke and speed of rotation can be altered independently of each other. The speed of rotation can be controlled by the rate of this flow and the stroke can be adjusted according to the pH of the input (Fig. 31).

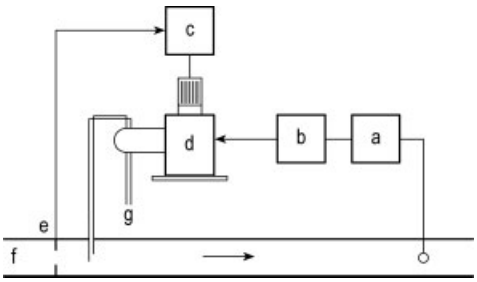

Figure 31. pH control with feedforward control and dosage pump
a) pH meter; b) Stroke regulator; c) Speed of rotation control; d) Dosage pump; e) Flow measurement; f) Mainstream, g) Additive

Cascade Control. When oscillations reach unacceptable amplitudes, a multistage series of controllers may be useful [121]. The first control loop is responsible for coarse control, where the major proportion of the additive is dosed. The purpose of the second controller is then to smooth out the control errors of the first controller by the addition of small volumes or dilute solutions. Each controller operates independently of the other. Two-step control can also be operated in such a manner that the first step always doses a slight excess and the fine control of the second step corrects this by controlled addition a unneutralized mainstream.

8. References

1. G. Bugge: *Der Alchemist – Die Geschichte Leonard Thurneyssers, des Goldmachers von Berlin,* W. Limpert-Verlag, Berlin 1943.
2. S. Arrhenius: "Über die Dissoziation der im Wasser gelösten Stoffe," *Z. Phys. Chem.* **1** (1887) 631 – 648.
3. W. Ostwald: *Grundlagen der Analytischen Chemie,* Th. Steinkopf-Verlag, Dresden 1894.
4. H. Friedenthal: "Die Bestimmung der Reaktion einer Flüssigkeit mit Hilfe von Indikatoren," *Z. Elektrochem.* **10** (1904) 113 – 119.
5. S. P. L. Sørensen: "Enzymstudien II. Über die Messung und die Bedeutung von Wasserstoffionen-Konzentrationen bei enzymatischen Prozessen," *Biochem. Z.* **21** (1909) 131 – 199.
6. H. Galster: *ph-Measurement, Fundamentals, Methods, Applications, Instrumentation,* VCH Verlagsgesellschaft, Weinheim 1991.
7. DIN 19267, pH-Messung, Technische Pufferlösungen vorzugsweise zur Eichung von technischen pH-Meßanlagen, 1978.
8. R. A. Robinson, R. H. Stokes: *Electrolyte Solutions,* Butterworths, London 1959, pp. 450 – 503.
9. P. Debey, E. Hückel: "Zur Theorie der Elektrolyte," *Phys. Z.* **24** (1923) 185 – 206.
10. R. G. Bates: *Determination of pH,* John Wiley, New York 1965.
11. G. G. Manov, R. G. Bates, W. J. Hamer, S. F. Acree: "Values of the Constants in the Debye-Hückel Equation for Activity Coefficients," *J. Am. Chem. Soc.* **65** (1943) 1765 – 1767.
12. J. Kielland: "Individual Activity Coefficients of Ions in Aqueous Solutions," *J. Am. Chem. Soc.* **59** (1937) 1675 – 1678.
13. K. Schwabe: *Physikalische Chemie,* **vol. 2,** Akademie Verlag, Berlin 1974, p. 65.
14. R. G. Bates, E. A. Guggenheim: "Report on the Standardisation of pH and Related Terminology," *Pure Appl. Chem.* **1** (1960) 163 – 169.

15. M. Eigen: “die ’unmessbar’ schnellen Reaktionen,” *Angew. Chem.* **80** (1968) 892 – 906.
16. G. J. Bingold, A. D. Brewer, B. Heam: “Specific Conductivity and Ionic Product of Water Between 50 and 271 °C,” *Trans. Faraday Soc.* **67** (1971) 2419 – 2430.
17. T. Mussini, F. Mazza: “Comparison of different pH Scales, Discussion Contribution,” *DECHEMA Monogr.* **101** (1986) 79 – 82.
18. S. Rondinini, P. Longhi, P. R. Mussini: “Autoprotolysis Constants in Non Aqueous Solvents and Aqueous Organic Solvent Mixtures,” *Pure Appl. Chem.* **59** (1987) 1692 – 1702.
19. B. B. Owen: “Direct Measurements of the Primary, Secondary and Total Medium Effects in Acetic Acid,” *J. Am. Chem. Soc.* **54** (1932) 1758 – 1769.
20. DIN 19261, pH-Messung, Begriffe für Meßverfahren mit Verwendung galvanischer Zellen, 1971.
21. W. Nernst: “Die elektrochemische Wirksamkeit der Ionen,” *Z. Phys. Chem.* **4** (1889) 129 – 181.
22. IEC 746-2, *Expression of performance of electro chemical analyzers,* Part 2: ph value, Bureau Central de la Commission Electrotechnique Internationale, Genève 1982.
23. M. Le Blanc: “Die elektromotorischen Kräfte der Polarisation II.,” *Z. Phys. Chem.* **12** (1993) 333 – 358.
24. D. J. G. Ives, G. J. Janz: *Reference Electrodes, Theory and Practice,* Academic Press, New York 1961.
25. R. G. Bates, V. E. Bower: “Standard Potential of the Silver – Silver – Chloride Electrode from 0 ° to 95 °C and the Thermodynamic Properties of Dilute Hydrochloric Acid Solutions,” *J. Res. NBS* **53** (1954) 283 – 290.
26. R. G. Bates, D. Rosenthal: ”Standard Potential of the Silver – Silver – Chloride Electrode and Activity Coefficients of Hydrochloric Acid in Aqueous Methanol (33.4 wt %) With and Without Added Sodium Chloride at 25 °,” *J. Phys. Chem.* **67** (1963) 1088 – 1090.
27. D. I. Hitchcock, A. C. Taylor: “The Standardization of Hydrogen Ion Determination II,” *J. Am. Chem. Soc.* **60** (1938) 2710 – 2717.
28. R. G. Bates: “Definitions of pH Scales,” *Chem. Rev.* **42** (1948) 1 – 61.
29. P. Spitzer: “Comparison of pH Values of Primary Buffer Materials,” *PTB-Bericht*, **PTB-W-68,** Braunschweig 1997.
30. DIN 19266, pH Messung, Referenzpufferlösungen zur Kalibrierung von pH-Meßeinrichtungen, 2000.
31. F. G. K. Baucke: “Differential-potentiometric Cell for the Restandardization of pH Reference Materials,” *J. Electroanal. Chem.* **368** (1994) 67 – 75.
32. A. K. Covington, R. G. Bates, R. A. Durst: “Definition of pH Scales, Standard Reference Values, Measurement of pH and Related Terminology,” *Pure Appl. Chem.* **57** (1985) 531 – 542.
33. R. G. Bates, M. Paalo, R. A. Robinson: “Interpretation of pH-Measurements in Alcohol – Water Solvents,” *J. Phys. Chem.* **67** (1963) 1833 – 1838.
34. W. Thomson: “Electrolytic Conduction in Solids – First Example: Hot Glass,” *Proc. Roy. Soc.* **A 23** (1875) 463 – 464.
35. M. Cremer: “Über die Ursache der elektromotorischen Eigenschaften der Gewebe, zugleich ein Beitrag zur Lehre von den polyphasischen Elekrolytketten,” *Z. Biol.* **46** (1906) 562 – 608.
36. F. Haber, Z. Klemenziewicz: “Über elektrische Phasengrenzkräfte,” *Z. Phys. Chem.* **67** (1909) 385 – 431.
37. R. Piosik, W. Jansen, R. Peter: “Die Erfindung der Glaselektrode durch Fritz Haber und Zygmont Klemensiewicz,” *Chem. Lab. Betr.* **44** (1993) 500 – 508, 573 – 575.
38. DIN 19263, pH Messung, Glaselektroden, 1989.
39. D. Hubbard: “Hygroscopicity and electrode function (pH Rasponsive) of Glasses as a Measure of Service Ability,” *J. Res. NBS* **36** (1946) 511 – 520.
40. H. Bach, F. G. K. Baucke: “Investigations of Glasses Using Surface Profiling by Spectrochemical Analysis of Sputter-induced Radiation I. Surface Profiling Technique with high In-Depth Resolution,” *J. Am. Chem. Soc.* **65** (1982) 527 – 533.
41. A. Wikby: “Chemical and Electrical Properties on the Surface Layers of some Glass Electrodes,” *Electrochim. Acta* **19** (1974) 329 – 336.
42. S. Dobos: “Strukturumwandlungen des Silikanetzwerkes im Laufe der Ausbildung der Oberflächenschicht bei natriumoxidhaltigen Gläsern,” *Acta Chim. Acad. Sci. Hung.* **69** (1971) 43 – 48.
43. F. G. K. Baucke, H. Bach: “Investigation of Glass Using Surface Profiling. II. Field driven

Formation and Electrochemical Properties of Protonated Glasses Containing Various Proton Concentrations," *J. Am. Ceram. Soc.* **65** (1982) 534 – 539.
44. K. Schwabe, H. Dahms: "Versuche zur Frage der Durchlässigkeit von Glaselektroden für Wasserstoffionen mit Hilfe von Tritium-Markierung," *Monatsber. Dtsch. Akad. Wiss.* **1** (1959) 279 – 282.
45. D. A. MacInnes, M. Dole: "The Behavior of Glass Electrodes of Different Compositions," *J. Am. Chem. Soc.* **52** (1930) 29 – 36.
46. G. A. Perley: "Glasses for Measurement of pH," *Anal. Chem.* **21** (1949) 394 – 401.
47. DIN 19264, pH-Messung, Bezugselektroden, 1985.
48. P. Mussini, F. D'Andrea, A. Galli, P. Longhi, S. Rondinini: "Characterization and Use of Aqueous Caesium Chloride as an Ultra-concentrated Salt Bridge," *J. Appl. Electrochem.* **20** (1990) 651 – 655.
49. F. G. K. Baucke: "Standardpotentiale ($\varepsilon^0 + \varepsilon_j$) der Silber/Silberchlorid-Elektrode in 3.5 m und ges. KCl unter Verwendung entsprechender ("Cl-ionenselektiver") Membranelektroden (= – 95 °C)," *Chem. Ing. Tech.* **47** (1975) 565 – 566.
50. H. Galster: "Natur, Messung und Anwendung der Redoxspannung, Teil I," *Chem. Lab. Betr. (CLB)* **30** (1979) 330 – 335.
51. W. Ingold: "pH-Meßgeber für Drücke bis 25 atü und Temperaturen über 100 °C," *DECHEMA Monog.* **43** (1962) 153 – 160.
52. H. K. Fricke: "Eine neue Bezugs- und Ableitelektrode für Gaselektroden-Meßketten," *DECHEMA-Monograph.* **43** (1962) 161 – 172.
53. C. H. Hamann, W. Vielstich: *Elektrochemie I,* Wiley-VCH, Weinheim 1998.
54. H. Galster: "Störungsquellen in der Bezugselektrode bei pH-Messungen," *Fresenius Z. Anal. Chem.* **245** (1969) 62 – 67.
55. Ingold Messtechnik, WO 92/21960, 1992 (A. Nipkow, E. Bakker).
56. K. Camman, H. Galster: *Das Arbeiten mit ionenselektiven Elektroden,* 3rd ed., Springer Verlag, Berlin 1996.
57. Ingold Messtechnik, CH 661 128, 1987 (J. Brinkmann, H. Bühler, A. Lorum).
58. H. Kaden, W. Vonau: "Bezugselektroden für elektrochemische Messungen," *J. prakt. Chem.* **340** (1998) 710 – 721.
59. Proton AG, DE 3 405 401 C2, 1985 (H. Bühler, H. Galster).
60. A. K. Covington, H. P. Büttikofer, M. G. C. Camos, M. I. A. Ferra, M. J. F. Rebelo: "Procedure for Testing pH Response Glass Electrodes at 25, 37, 65 and 85 °C. Determination of Alkaline Errors up to 1 mol dm^{-3} Na^+, K^+, Li^+," *Pure Appl. Chem.* **57** (1985) 887 – 898.
61. B. Emmerich: "Die Emailelektrode, eine neue Lösung zur pH-Messung im Betrieb," *Regelungstech. Praxis* **20** (1978) 313 – 317.
62. K. Schwabe: *pH Messtechnik,* Th. Steinkopf, Dresden 1976.
63. DIN 19265, pH-Messung, pH-Meßumformer, Anforderungen, 1994.
64. L. Micic, A. Gilly: "Gleichspannungsverstärker mit Kapazitätsdioden für kleine Eingangsleitungen," *Electronic* **12** (1963) 263 – 265.
65. L. Wilhelmy: "Der Trennverstärker, ein neues Bauelement zur pH-Messung," *Meßtechnik* **1973** 87 – 89.
66. P. Wunderer: "pH Meßumformer und angeschlossene Sichtausgeber" in *Richtlinien für die pH Messung in industriellen Anlagen, Normenheft 22,* Beuth-Verlag, Berlin 1974.
67. G. Clifford: "Collecting pH and Specific Ion Measurements with a PC," *Am. Lab.* **19** (1987) 62 – 68.
68. Knick, DE 19 536 315 A1, 1997 (W. Feucht).
69. Mettler-Toledo, EP 0 419 769 A2, 1991 (J. Gareis).
70. Knick, EP 0 497 994 A1, 1992 (T. Martell).
71. Knick, DE 19 539 763 A1, 1997 (W. Feucht).
72. Knick, DE 2 937 227 C2, 1980 (W. Feucht).
73. WTW, DE 3 505 342 C2, 1986 (R. Keller).
74. W. Melzer: "Automatisierte pH-Messung mit On-line-Steilheitskompensation der Glaselektrodenmeßkette," *Tech. Mess.* **61** (1994) 208 – 214.
75. Schott-Geräte GmbH, DE 4 438 524 A1, 1996 (H. Diehm et al.).
76. DIN 19 268, pH-Messung von klaren, wäßrigen Lösungen, 1985.
77. F. G. K. Baucke, R. Naumann, C. A. Weber: "Multiple-Point Calibration with Linear Regression as a Proposed Standardization Procedure for High-Precision pH Measurements," *Anal. Chem.* **65** (1993) 3244 – 3251.
78. S. Ebel, E. Glaser, H. Mohr: "Fehler und Fehlerfortpflanzung bei der Bestimmung von pH Werten," *Fresenius Z. Anal. Chem.* **293** (1978) 33 – 35.

79. S. Ebel: "Linearisierte und nichtlinearisierte Regression bei der Ermittlung von Kalibrierkurven," *GIT* **4** (1984) Suppl., 6 – 8.
80. G. K. McMillan: "Understand Some Basic Truth of pH Measurement," *Chem. Ing. Prgr.* (1991) Okt., 30 – 37.
81. DIN 38 404, Teil 5: Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Physikalische und physikalisch-chemische Kenngrößen. (Gruppe C). Bestimmung des pH-Wertes (C5), 1984.
82. ISO 10523, International Standard. "Water quality – Determination of pH," 1994 (E).
83. Der Bundesminister für Jugend, Familie und Gesundheit (ed.): *Verordnung über Trinkwasser und Lebensmittelbetriebe,* 22. Mai 1986, part 1, pp. 760 – 773.
84. DIN 38 404, Teil 10: Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Physikalische und physikalisch-chemische Kenngrößen. (Gruppe C). Calciumcarbonat-Sättigung eines Wassers. 1990.
85. VDI-Richtlinien VDI 3870, Part 10. Analysis of Rainwater: Determination of the pH Value of Rainwater, 1994.
86. H. Galster: "pH-Messung in vollentsalztem Wasser," *Kraftwerkstechnik* **59** (1979) 885 – 889.
87. W. F. Koch, G. Marienko, R. C. Paule: "Development of a Standard Reference Material for Rainwater Analysis," *J. Res. NBS* **91** (1986) 33 – 41.
88. H. Hammerschmidt, J. Putzien, J. Winkler: "Kontinuierliche pH-Messung im Einlauf der Kläranlage," *Korrespondenz Abwasser* **34** (1987) 730 – 733.
89. M. A. Clarke: "The Effect of Solution Structure on Electrode Processes in Sugar Solutions," *Proc. of the 1970 Technical Session on Cane Sugar Refining Res.* 12th – 13th Oct. (1970) 179 – 188.
90. A. Emmerich, C. Reichel: "Über die pH-Bestimmung in Rübenzucker," *Zucker* **14** (1961) 340 – 343.
91. K. Schwabe: "Azidität konzentrierter Electrolytlösungen," *Electrochim. Acta.* **12** (1967) 67 – 93.
92. K. Schwabe: "Elektrochemische pH-Messung unter extremen Bedingungen," Monogr. zu Angew. Chem. and Chem. Ing. Tech., Weinheim 1960, 26 – 29.
93. S. Licht: "pH Measurement in Concentrated Alkaline Solutions," *Anal. Chem.* **57** (1985) 514 – 519.
94. G. Wiegner: "Über Wasserstoff- und Hydroxylionen in Ionenschwärmen um suspendierte Teilchen und suspendierte Ultramikronen," *Kolloid Z.* **51** (1930) 49 – 60.
95. H. Pallmann: "Die Wasserstoffaktivität in Suspensionen und kolloiddispersen Systemen," *Kolloidchem. Beih.* **30** (1930) 334 – 405.
96. L. Milicka: "Zum Problem der pH-Messung in Suspensionen," *Z. phys. Chem.* **249** (1972) 63 – 72, 117 – 180.
97. H. Jenney, T. R. Nielsen, N. T. Coleman, D. E. Williams: "Concerning the Measurement of pH, Ion Activities and Membrane Potentials in Colloidal System," *Science* **112** (1950) 164 – 167.
98. D. P. Brezinski: "Influence of Colloidal Charge on Response of pH and Reference Electrodes: The Suspension Effect," *Talanta* **30** (1983) 347 – 354.
99. K. Spiegel, W. Möhler: "pH Messung in der Verfahrenstechnik," *DECHEMA Monogr.* **126** (1992) 373 – 388.
100. W. M. Baumann, W. Simon: "Säure-Base-Titrationen in flüssigem Ammoniak," *Fresenius Z. Anal. Chem.* **216** (1966) 273 – 278.
101. K. Cruse: "pH-Begriff und pH-Messung in wasserfreien Lösungen," *Arch. Tech. Mess.* (1956) 125 – 128, 169 – 172, 203 – 204, 217 – 220.
102. R. G. Bates, M. Paabo, R. A. Robinson: "Interpretation of pH-Measurements in Alcohol-Water Solvents," *J. Phys. Chem.* **67** (1963) 1833 – 1838.
103. T. Mussini, A. K. Covington, P. Longhi, S. Rondini: Criteria of Standardisation of pH Measurements in Organic Solvents and Water-Organic Solvent Mixtures of Moderate to High Permittives," *Pure Appl. Chem.* **57** (1985) 865 – 876.
104. R. Degner, St. Leibl: *pH messen, So wird's gemacht!* VCH Verlagsgesellschaft, Weinheim 1995.
105. L. Van den Berg: "pH Measurement at low Temperatures Using Modified Calomel and Glass Electrodes," *Anal. Chem.* **32** (1960) 628 – 631.
106. D. Midgley: "A review of pH measurement at High Temperatures," *Talanta* **37** (1990) 767 – 781.
107. S. Hettiarachchi, K. Makela, H. Song, D. D. MacDonald: "The Viability of pH measurements in Supercritical Aqueous Systems," *J. Electrochem. Soc.* **139** (1992) L3 – L4.

108. A. Fiechter, W. Ingold, A. Baerfuß: "Die pH-Kontrolle in der mikrobiologischen Verfahrenstechnik," *Chem. Ing. Tech.* **36** (1964) 1000 – 1004.
109. H. Galster: "Schutz von pH-Meßanlagen vor elektromagnetischen Feldern," *Tech. Mess.* **49** (1982) 189 – 192.
110. WTW, DE 3 518 870 A1, 1986 (G. Gügel).
111. Berliner Entwässerungswerke, Bericht: FEV Wasser 102 04 318 (1987) Teilvorhaben pH-Wert.
112. Siemens AG, DE 1 804 962, 1974 (E. Michel, K. H. Papst).
113. Proton AG, DE 2 061 978, 1972 (W. Ingold).
114. Conducta GmbH, DE 3 927 282, 1991 (K. Stellmacher, K. H. Oppermann).
115. N. Schulte: "Die prozesstechnische pH-Messung," *Chem. Anl. Verf.* **1993** H.3. 29 – 34.
116. Ingold AG, CH 673 783, 1990 (U. Schaad, J. Brinkmann).
117. E. G. Shinskey: *pH and pIon Control in Process and Waste Streams,* John Wiley, New York 1973.
118. O. Künzel: "Regelkreise mit Zweipunktreglern," *Elektrotechnik* **53** (1971) no. 1/2, 18 – 21.
119. H. Kubli: "Ein einfaches Präzisions-pH-Regelsystem für die mikrobiologische Verfahrenstechnik," *Int. Jahrb. Chem. Ind.* **1967,** 3 – 4.
120. E. Samal: "Verbesserung der Regeltüte durch Störgrößenaufschaltung," *Regelungstech.* **5** (1957) 40 – 45, 154 – 158.
121. E. Samal: "Verbesserung der Regeltüte durch Verwendung von Hilfsgrößenbeeinflussung oder Reihenschaltung mehrerer Regler," *Regelungstech.* **5** (1957) 192 – 198.

Quasicrystals

Enrique Maciá, GISC Universidad Complutense de Madrid, Madrid, Spain

Jean-Marie Dubois, Ecole des Mines, Nancy, France

Patricia Ann Thiel, Iowa State UniversityAmes, Iowa 50011, United States

1. Introduction 1
2. Structural Aspects 2
3. Preparation of Quasicrystalline Phases 3
4. Chemical Composition 3
5. Electrical Conductivity 4
6. Thermal Conductivity and Diffusivity . 7
7. Mechanical and Surface Properties . . . 8
8. Potential Applications 10
9. References 12

1. Introduction

Metals and their alloys have played an important role in the history of human civilization. This is conveniently illustrated by the fact that two major stages of civilization are known as the Iron Age and the Bronze Age. This helps to place in context the fundamental importance of the discovery, starting in 1982, of a new family of metallic alloys that triggered a deep revision of our understanding of the structure of matter [1–6]. This revision ultimately led to the introduction of the concept of *order without periodicity* (aperiodic order) and, at the same time, spurred intense interdisciplinary activity aimed at a deeper understanding of natural systems in which aperiodic order spontaneously arises.

To understand the conceptual revolution ignited by the discovery of these alloys, it is necessary to recall that the traditional classification scheme of solid matter is based on the notion of *periodic ordering* of atoms in space. It includes two wide categories: crystalline matter, in which the regular location of the constituent atoms determines a long-range order exhibiting translational symmetry; and amorphous matter, in which, although there are correlations in the positions of neighboring atoms, long-range order is completely absent. The implicit dichotomy in this classification scheme made crystalline matter the archetype of *order* in solid-state physics, while amorphous matter represented the paradigm of *disorder*.

This traditional scheme was shaken by the claim that the new alloys, although not conventionally crystalline, nonetheless exhibited discrete, high-quality diffraction patterns. In other words, they could not be regarded as amorphous, since their constituent elements showed a remarkable long-range ordering. The name *quasicrystal* rapidly became popular for this kind of material [7].

From this discovery, the following picture eventually emerged: In quasicrystals, atoms occupy space according to a *quasiperiodic* arrangement instead of the usual periodic fashion observed in crystals. A simple understanding of what a quasiperiodic arrangement might be can be gained by thinking of the Fibonacci sequence. There, the nth number is entirely predictable (the sequence is "ordered") but the rule used to generate it is not analogous to periodicity: each number is the sum of the preceding two, that is, $n_j = n_{j-2} + n_{j-1}$. (This is in contrast to an arithmetic sequence, in which the rule of generation *is* analogous to periodicity.) The geometrical analogue of the Fibonacci sequence constructed by using two types of segments is thought to represent a "one-dimensional quasicrystal", and real quasicrystals indeed often exhibit features which can be described successfully in terms of the Fibonacci sequence.

But in addition to their importance in fundamental research, the interest aroused by these new materials encompassed the field of technological innovations. In fact, the first industrial patent for a straightforward application of quasicrystals was obtained in 1988 [8]. Since then, the number of patents related to quasicrystalline phases has progressively increased and, at present, quasicrystal-containing alloys are pro-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.e22_e01

duced at a rate of at least several tonnes per annum.

In the years since their discovery, quasicrystalline phases have been observed in many more alloys, including binary, ternary, and quaternary compounds, mainly based on aluminum and titanium. In addition to the fivefold symmetry axes characteristic of the icosahedral group, eightfold, tenfold, and twelvefold symmetries have been discovered, all of which are incompatible with periodic translational symmetry [3–5]. Figure 1 shows the approximate distribution of the quasicrystalline phases observed to date, according to their symmetries. The overwhelming majority belong to the icosahedral group.

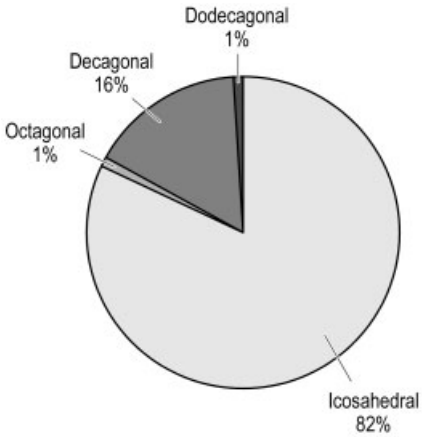

Figure 1. Percentage distribution of the four major symmetry types of quasicrystalline phases obtained to date Icosahedral (i) phases exhibit fivefold symmetry, octagonal (o) phases eightfold symmetry, decagonal (d) phases tenfold symmetry, and dodecagonal (dd) phases twelvefold symmetry

Due to the increasing number and variety of known quasicrystals, the International Crystallographic Union has redefined the term *crystal* to include this new kind of order. In this definition, a crystal is any solid which possesses an essentially discrete diffraction pattern, thus transferring the essential attribute of crystallinity from real space to reciprocal space. Consequently, within the crystalline family we can distinguish between *periodic crystals* and *aperiodic crystals*. This wider definition reflects the present idea that microscopic periodicity is a sufficient but not necessary condition for crystallinity. In this sense, it is interesting to note that the first mention ever made about using the concept of aperiodic crystals in solid-state physics was probably due to SCHRÖDINGER [9], who introduced it in his 1944 discussion on the possible nature of genetic material. Finally, an interesting aspect is the possible relationship between quasicrystalline and conventional crystals, in particular the possible structural transition from one kind to the other, an aspect that has received a great deal of attention.

2. Structural Aspects

A feature of quasicrystalline order is the *self-similarity* of the structure, which constitutes an essential point in most of the structural models proposed for these materials to date [10–14]. This self-similarity and, more specifically, the scale on which features appear self-similar, is referred to as the *inflation symmetry*.

This self-similarity is illustrated in Figure 2, which shows the fivefold arrangement of spots characteristic of the diffraction pattern of an icosahedral quasicrystal and a geometric construction which bears similar properties. (Under many diffraction conditions, a fivefold symmetry like that shown in Figure 2 a appears tenfold because of the introduction of a center of inversion symmetry.) In Figure 2 b, the points of intersection of the lines that join the pentagon vertices define a new pentagon. Its size has a definite relationship to that of the original one, determined by the ratio τ^{-2}, where τ is the ratio between the diagonal and the side of the pentagon. This ratio, corresponding to the irrational number $(\sqrt{5}+1)/2$, has been known since ancient times as the *Golden Mean*, and it is implemented in arts such as painting, sculpture, and architecture. The Golden Mean also describes fundamental properties of the Fibonacci sequence. It is remarkable that the geometric construction shown in Figure 2 b, known as a Pitagoric pentagram, is naturally expressed by the spatial atomic arrangement within a quasicrystal, as can be readily checked by means of a careful comparison with a real diffraction pattern, such as that shown in Figure 2 a. Another remarkable feature of quasicrystalline structures are that they can be "simplified" by lifting them into a higher-dimensional space (higher than the

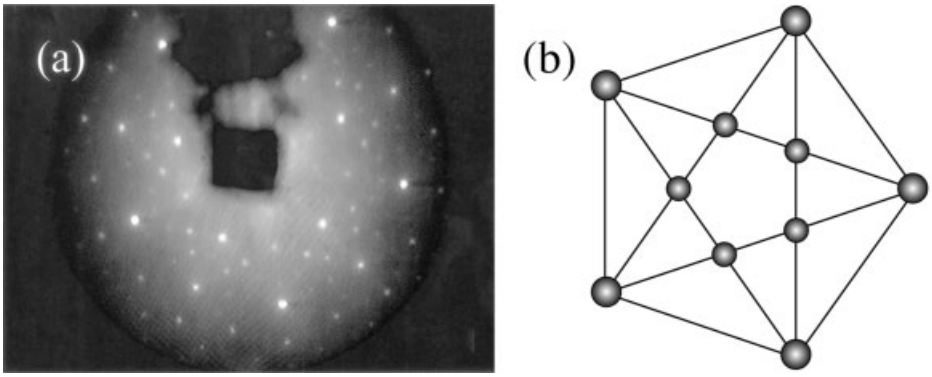

Figure 2. a) Low-energy electron diffraction pattern of the fivefold surface of an icosahedral phase of an Al – Pd – Mn quasicrystal, in which the existence of pentagonal patterns associated with fivefold axes can be clearly seen [18, 19]. b) Diagram illustrating the self-similarity of the characteristic diffraction pattern of an icosahedral quasicrystal along a fivefold zone axis.

physical real space in which quasiperiodicity is manifested) [15–17].

Clusters containing several tens of atoms are also frequently mentioned in the discussion of quasicrystallographic structure. In particular, Mackay-like clusters and Bergman clusters have been identified in a number of alloys. It has been proposed that these clusters possess a special stability, that they constitute the fundamental building blocks of the solid, and that they are ultimately responsible for all of the unusual physical properties of quasicrystals. This hypothesis remains under active consideration. Figure 3 a shows the structure of one such cluster which may exist in icosahedral Al – Pd – Mn, and Fig 3 shows an arrangement of these clusters along the fivefold axis. Note that the clusters exhibit the same type of inflation symmetry as is illustrated in Figure 2 b.

Atomic quasiperiodic order is not just limited to the bulk of the solid, but is present also on its surface. This is manifest, for instance, in the growth habits of monograins, which in icosahedral systems often display dodecahedral or triacontahedral shapes [20], as predicted by theory [21, 22]. In decagonal systems, the growth habits are typically more anisotropic, yet still can reflect the quasicrystalline nature. Examples are shown in Figure 4 [23–25].

3. Preparation of Quasicrystalline Phases

While many quasicrystalline phases are metastable, a number of thermodynamically stable phases also exist, including Al – Cu – Li, Al – Cu – (Fe, Ru, Os) and Al – Pd – (Mn, Re) [26]. The latter materials preserve a quasiperiodic structure of extraordinary quality until they melt, which occurs at about 1100 – 1200 K. A detailed knowledge of the phase diagrams of such alloys and the corresponding ability to precisely determine the boundaries of the small stability regions of the different quasicrystalline phases have made it possible to grow high-quality single-grain quasicrystals by the conventional Bridgman and Czochralski growth techniques, and by the less-conventional technique of flux growth (slow cooling) [26]. There has been considerable progress in the size and quality of samples available. Optimal single-grain sizes have increased from a few micrometers at their discovery, to about one centimeter a decade later, to about ten centimeters at present.

The discovery of the thermodynamically stable phases also enabled other conventional techniques to be used for preparation of quasicrystals in multiple forms, e.g., gas and centrifugal atomization for preparation of powders; magnetron sputtering for thin films; thermal spray techniques for coatings; and sintering and pressing for solid bodies, to name but a few.

4. Chemical Composition

In Figure 5, the various chemical elements which constitute the different known quasicrystalline phases are indicated via a histogram illustrating their relative frequency of appearance.

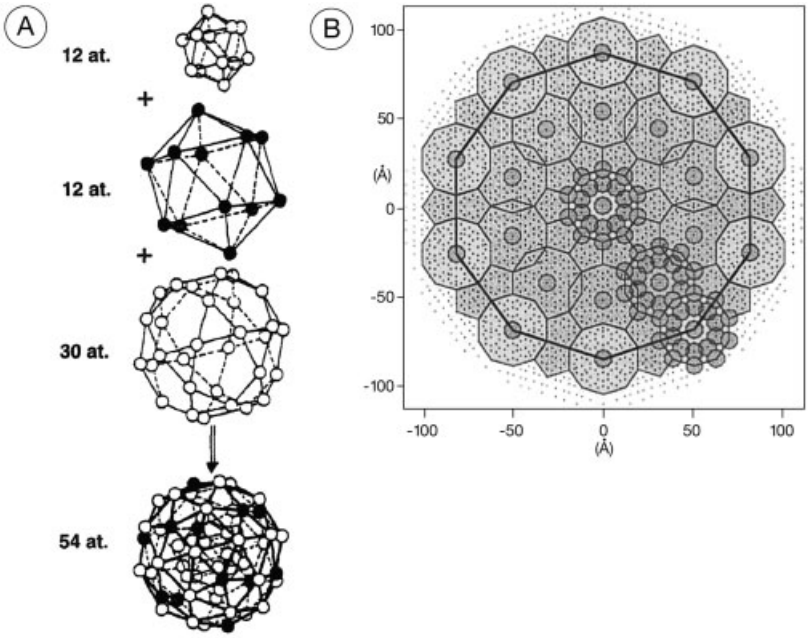

Figure 3. a) Structure of a Mackay icosahedron. Similar clusters can be identified in the bulk of icosahedral Al – Pd – Mn and Al – Cu – Fe, although other types of clusters can be identified as well. b) Arrangement of Mackay-type icosahedra showing the hierarchical, self-similar arrangement of overlapping clusters [11, 13, 14].

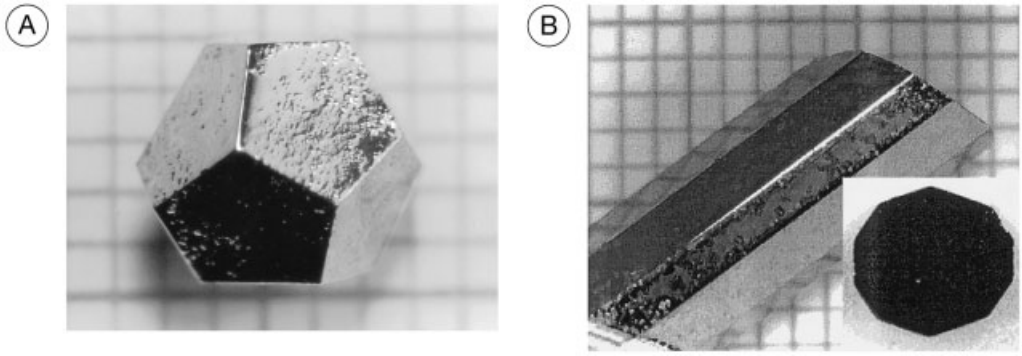

Figure 4. Single grains of quasicrystals resting on a 1×1 mm grid
A) A single grain of Ho – Mg – Zn. The morphology is that of a pentagonal dodecahedron, consistent with the icosahedral symmetry of the lattice. Photograph courtesy of I. R. Fisher and P. C. Canfield, Ames Laboratory [25]. B) A single grain of decagonal Al – Ni – Co. The long axis of this sample is parallel to the tenfold axis, which is perpendicular to the decagonal (aperiodic) planes. The sample has ten long facetted sides, five of which are visible in the main photograph. The inset shows the face of a similar crystal cut perpendicular to the tenfold axis, thereby revealing all ten facets. Photograph courtesy of I. R. Fisher, N. D. Kelso, and P. C. Canfield, Ames Laboratory.

Aluminum is an almost ubiquitous atomic species in quasicrystals, since it is present in about 85 % of the phases. Other elements which are major constituents of certain quasicrystalline phases include titanium, cadmium, zinc, magnesium, and gallium [26].

The composition of quasicrystalline alloys is quite variable, although it is generally thought to fulfill particular electron-to-atom ratios, similar to the requirement for Hume – Rothery phases [27–30]. The following serve to illustrate some of the known compounds: $Al_{86}Mn_{14}$, $Al_{65}Cu_{20}Fe_{15}$, $Al_{65}Co_{20}Cu_{15}$, $Al_{70}Pd_{21}Mn_{9}$, $Zn_{60}Mg_{30}RE_{10}$ (RE = Y, Dy, Gd, Ho, Tb, Er), and $Ti_{45}Zr_{38}Ni_{17}$. Most known stable phases contain three or more elements, but several

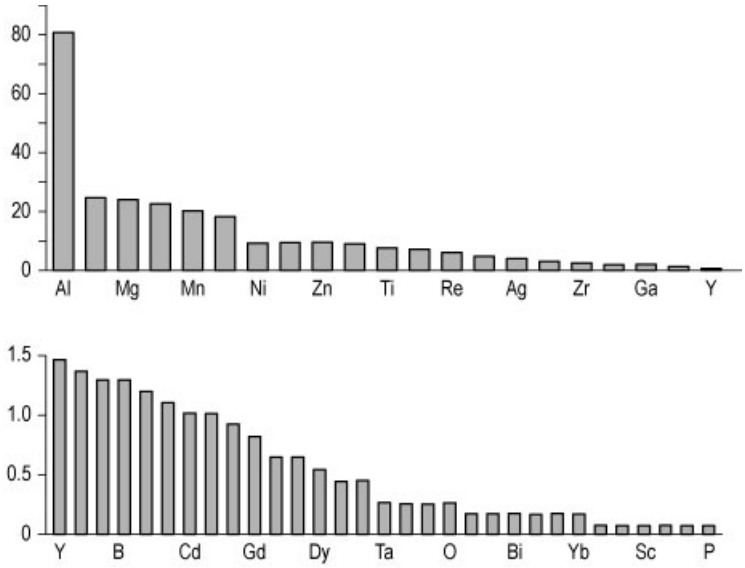

Figure 5. Frequency of appearance of different elements present in the various quasicrystalline alloys obtained to date

two-component systems are notable exceptions: $Cd_{5.7}Yb$ and $Cd_{5.7}Ca$ (both icosahedral) [31] and $Ta_{1.6}Te$ (dodecagonal) [32]. In terms of chemical composition, the largest family consists of alloys with about 60 – 75 atom % Al.

The compositional range which supports a given thermodynamically stable phase is typically very small, spanning only a few atom percent. (Because of this, quasicrystals can be appropriately termed intermetallic compounds.) However, related crystalline phases called *approximants* usually surround the quasicrystalline region in the phase diagrams. The approximants bear strong structural resemblances to the quasicrystals, and their properties are often similar as well. In terms of properties, the quasicrystalline phase typically exhibits a local minimum or maximum value, but the adjacent approximants form a bridge to the ordinary crystalline phases. For some technical applications, the approximants may be even more appealing than the quasicrystals, since they are less sensitive to small compositional changes.

5. Electrical Conductivity

Once the existence of a new kind of ordered matter has been established, it seems natural to wonder to what extent the new structural properties of such materials can determine their transport properties. In the field of quasicrystals, the most widespread opinion is that quasiperiodic long-range order should have a significant impact on most physical properties. And, in fact, from the diverse measurements carried out to date of the electrical resistivity, thermoelectric effect, magnetoresistivity, Hall effect, specific heat, magnetic susceptibility, and optical properties, it appears that the transport properties of quasicrystals are, in general, very different from those expected for a simple combination of the constituent metals [33, 34].

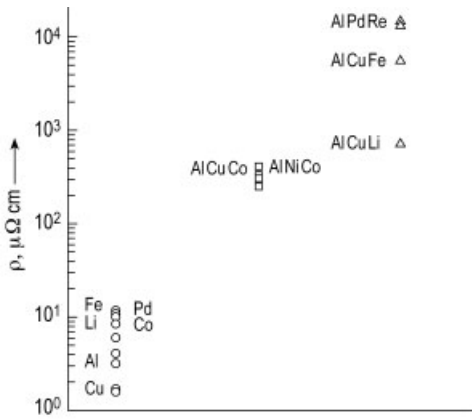

Figure 6. Comparison of the electrical resistivities of some stable quasicrystalline phases and their constituent elements

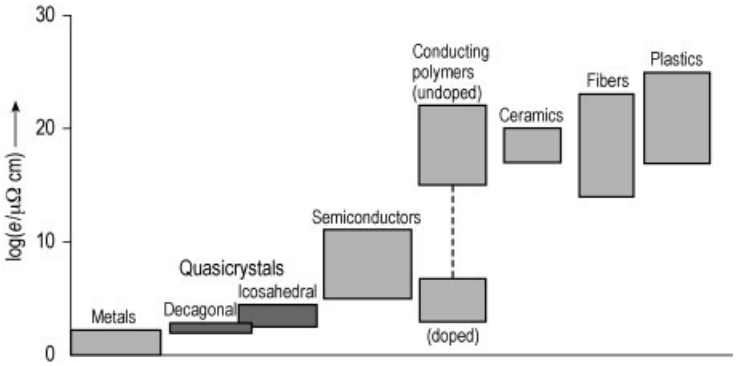

Figure 7. Comparison of the electrical resistivities of quasicrystals and other broad classes of materials of technological interest

This important result is illustrated in Figure 6, which gives the electrical resistivities ρ measured at room temperature for quasicrystals and their constituent (pure) elements. It can be seen that the isolated elements are good metals possessing low electrical resistivities. In contrast, quasicrystalline alloys comprised of these elements show electrical resistivities which are higher by several orders of magnitude, and much higher than typical values for conventional metallic alloys (both crystalline and amorphous), whose representative values fall in the region between the horizontal dashed lines in Figure 6.

In addition, unlike metals, the electrical conductivity $\sigma = \rho^{-1}$ of quasicrystals *increases* with increasing temperature, instead of decreasing, following a power law of the form $\sigma \propto T^{\alpha}$, where $1/2 \leq \alpha \leq 3/2$. Also in an anomalous manner, the electrical conductivity *decreases* when the quasiperiodic order of the structure is improved, for instance, by means of thermal treatment. And, to complete the list of oddities, electrical conductivity of quasicrystals depends sensitively on the chemical composition of the alloy. This has made it possible to observe a metal–insulator transition for the quaternary icosahedral alloy Al – Pd – Mn – Re, when the Mn content in the alloy is changed by a mere 5 %.

To summarize, thermodynamically stable quasicrystalline phases are clearly beyond the realm of metals in terms of their electrical transport properties, and in fact their inclusion in the category of semiconductor materials has been suggested [35]. In this sense, it is worth mentioning the analogy between the sensitive dependence of conductivity on chemical composition observed, for instance, for the icosahedral $Al_{62.5}Cu_{25}Fe_{12.5}$ alloy, for which an increment of 0.5 % in Al content is enough to double the conductivity value at low temperature, and the characteristic behavior of conductivity in doped semiconductors. In this regard, Figure 7 compares the electrical conductivities of several broad classes of materials.

As with many properties, the effect of quasicrystalline order on transport properties could be revealed best by comparing properties of samples of similar composition but different structure. The discovery of stable phases with decagonal symmetry made it possible to perform these measurements for the same sample, since the decagonal quasicrystals possess periodic order along one direction and quasiperiodic order within the perpendicular plane. In this way, it was confirmed that the electrical resistivity in the decagonal alloys Al – Cu – Co and Al – Ni – Co is high and first increases, then decreases with temperature along the quasiperiodic plane. In contrast, along the periodic axis the electrical resistivity is lower, and increases with increasing temperature, as is typical for crystalline metals [36]. This is illustrated in Figure 8 [24]. Thus, the marked anisotropy in the transport properties must be a consequence of the anisotropy in the sample's structural order.

Why do aperiodic systems exhibit anomalous electrical transport properties? Realistic models must encompass two factors. First, is the marked decrease in the number of charge carriers able to effectively contribute to the electrical conductivity. This decrease, ultimately due

to structural stabilization, is called a *pseudogap* in the density of electronic states. Its existence has been confirmed by means of many experimental techniques [37]. Second is the lack of translational symmetry in these systems (in two or three dimensions), which precludes the usual understanding of extended electronic states in a periodic lattice based on Bloch functions [38].

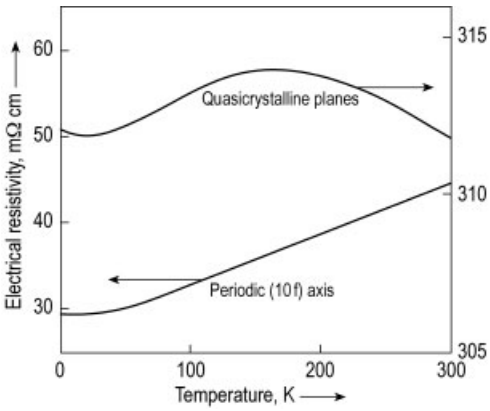

Figure 8. Anisotropy in the electrical resistivity of decagonal phases [24]

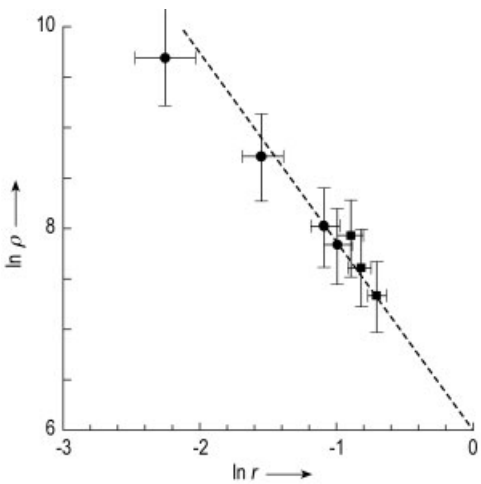

Figure 9. Resistivity at 300 K versus r. The parameter r is defined in the text. The straight line has a slope of -2. Circles represent icosahedral quasicrystals (top to bottom: Al – Pd – Re, Al – Pd – Mn, Al – Cu – Ru, and Al – Cu – Fe), and squares represent approximants in the Al – Cu – Fe system
Courtesy of Jean Marie Dubois, Ecole des Mines, Nancy, France

The real importance of the first factor, the pseudogap, is illustrated in Figure 9, which shows the resistivity ρ at 300 K as a function of a parameter r, which is defined [39] as the ratio of the density of states at the Fermi edge to the corresponding free-electron value, that is, r is a direct measure of the depth of the pseudogap. It turns out that the quasicrystals and their approximants exhibit a relationship $\rho \sim r^{-2}$ at constant temperature, which is obeyed also for disordered systems [40]. While the pseudogap is certainly important, it is not sufficient to explain all the transport phenomena, in particular the anomalous temperature dependence of the conductivity noted above.

6. Thermal Conductivity and Diffusivity

It is well known that most metals are good conductors of heat, and that thermal conductivity is proportional to the product of their electrical conductivity and temperature. None of this holds true for quasicrystals, which are very bad conductors of heat, basically because the density of free electrons is low. Consequently, heat must propagate by means of atomic vibrations (phonons). Phonons, which are periodic waves, do not travel easily in an aperiodic lattice once their wavelength becomes comparable to the size of the constituent atomic clusters, that is, above about 70 K. Furthermore, the vectorial combination of two phonons travels in the opposite direction [an Umklapp process; German: *umklappen* = fold back)) to that of the incident phonons in quasicrystals, in strong contrast to the case of regular crystals, in which such a combination is favorable to heat transport. Finally, quasicrystals contain a large number of atomic sites with two nearly equivalent equilibrium positions. Such so-called phason sites may be excited by the traveling wave and hence absorb part of its energy. This also inhibits the conduction of heat.

Some representative values for thermal conductivities (in $\mathrm{W\,m^{-1}\,K^{-1}}$) of quasicrystals and other materials of industrial interest are listed in the following:

Ag	420
Cu	380
Al	209
Fe	59
GaAs	40
SiC	35 – 155
TiC	33
TiN	24
Al_2O_3	27 – 43
Decagonal quasicrystals	1.5 – 2.5
Icosahedral quasicrystals	2
Yttria-doped zirconia	2
Pyrex	1.3

The thermal conductivities of quasicrystals, both decagonal and icosahedral, are clearly lower than those of thermal insulators that are extensively used in aeronautical industry, such as titanium carbides or nitrides, doped zirconia, or alumina. In addition, the thermal diffusivity of these new materials is extraordinarily low, even lower than that of zirconium oxide. And, remarkably, the value of this diffusivity increases only by a factor of two with increasing temperature until the material melts.

7. Mechanical and Surface Properties

Table 1 compares mechanical and surface properties of quasicrystals with those of more common materials. The focus on the quasicrystalline phases in the Al – Pd – Mn and Al – Cu – Fe systems is because these data are more abundant than for other quasicrystalline alloys. The quasicrystals are noteworthy for their hardness, low surface energy, and low friction. An unattractive property is brittleness, as gauged by their low fracture toughness. Other attractive properties, for which data are not shown, include oxidation resistance, hydrogen storage capacity, and transition from brittle to ductile behavior at about three-fourths of the melting point.

Quasicrystals are very brittle. The toughness of single grains is less than 0.5 MPa $m^{1/2}$. In contrast, metallic crystals like aluminum are characterized by toughnesses about two orders of magnitude higher. As a result quasicrystals have little resistance to crack propagation below 450 °C. Hence, they are not very useful as bulk materials but may be more fruitfully employed as surface coatings or as composites. In such a case, the mechanical integrity is supplied by the substrate or the matrix, while the quasicrystal offers supplemental functions. For instance, in a coating, this may be hardness, protection against corrosion, optical absorbance, etc.

However, the brittleness of quasicrystals is overcome at high temperatures. Above about 450 °C, quasicrystals become plastic and deform more and more with applied stress. Deformation is facilitated by dislocations, as in conventional crystals. However, in contrast to conventional crystals, there is no work hardening. This proves the absence of pinning centers and indicates a viscous flow regime that might be due to some internal friction between supposedly undeformable atomic entities. Current research seeks actively to understand the plasticity of quasicrystals. In the same temperature range, it is also observed that thermal vacancies multiply and play the key role in atomic transport over large distances. Accordingly, diffusion coefficients of quasicrystals in the temperature range above 450 °C are similar to those measured in aluminum-based intermetallics. Similarly, the oxidation resistance of quasicrystals does not significantly depart from that of aluminum alloys or compounds. When in contact with pure O_2 or air, the quasicrystal surface is soon covered by a thin layer of pure alumina. Again, two temperature regimes may be distinguished, below and above 450 °C, with slow and quick formation of the oxide film, respectively.

Easy crack-propagation brittleness is indicative of a low cleavage energy (the energy required to split a crystal into two parts with no plastic deformation; for instance, at very low temperature). This energy is equal to twice the surface energy of the material, since two fresh surfaces are created by propagating the crack. In good metals like aluminum, the energy of the clean-surface is typically 1 J/m^2. No investigation of the clean surface energy has yet been performed on quasicrystals. Nevertheless, indirect evidence that the surface energy of quasicrystals is indeed small may be deduced from a comparison of contact angles measured by depositing small liquid droplets on the surface of a quasicrystalline sample as well as on reference samples such as teflon, metals, or a bulk oxide such as alumina. By using the approach developed two centuries ago by Young, and making certain assumptions, the contact angle can be related fairly simply to the reversible adhesion

Table 1. Comparison of the properties of quasicrystals with those of common metals and nonmetals

Material	Density, g/cm^3	Hardness, HV	Fracture toughness, MPa m$^{1/2}$	Young's modulus, GPa	Coefficient of friction ∗	Adhesion energy with water, mJ/m^2
Quasicrystals						
Icosahedral Al – Pd – Fe (single grain)		700 – 900	0.3	200	0.055	72
Icosahedral Al – Cu – Fe (polygrain)	4.6	800 – 1000	1.6	68 – 85	0.046	55
Common metals						
Copper			30	130		105
Aluminum	2.9	25	40	70	0.112	98
High-carbon steel		881			0.071	98
Low-carbon steel	7.8	120		215		
Nonmetals						
Yttria-doped zirconia	5.6					
Alumina		1950	4.7	300	0.067	120
Diamond		>5000			0.04 – 0.05	

∗ Unlubricated, with respect to a diamond pin in air.

energy W_L of a given liquid onto the surface, larger angles and poorer wetting corresponding to smaller values of W_L.

Poor wetting is an appropriate description of the interaction between carefully polished Al-rich quasicrystals and water. Correspondingly, their reversible adhesion energy with water is only 25 % larger than that of teflon (55 and 44 mJ/m^2, respectively) and one-third of that of window glass. The fundamental origin of this behavior is not yet understood, but a natural question is whether it is controlled by the properties of the surface of the quasicrystal, or those of the bulk. Quasicrystals always carry a thin surface oxide layer, consisting mainly of aluminum oxide. However, bulk alumina is comparable to window glass in that it is wet completely by water, indicating a high value of W_L. Thus, wetting of the quasicrystals is probably not controlled, at least to zeroth order, by the surface layer, but rather in some way by the underlying atomic and electronic structure of the bulk. (Note that this discussion is only demonstrably valid for water; common organic solvents wet quasicrystals well, and this indicates that the wetting may depend on the polarity of the liquid.)

Surface friction is a further important issue for quasicrystals. As shown in Table 1, numerous experiments using a spherical diamond pin have demonstrated that friction is lower on a quasicrystal ($\mu = 0.04 - 0.05$) than on steel ($\mu = 0.07$) of comparable hardness. With more conventional riders such as mild steel or WC – Co, the comparison is less favorable due to the transfer of material from the indenter to the surface. During such friction measurements, a major drawback arises from the brittleness of the contacting bodies and hence formation of wear particles. However, scratch tracks indicate that quasicrystals subjected to repetitive and severe shear develop some ductility during this process, and consequently acquire an ability to self-repair. This brittle-to-ductile behavior might be associated with the nucleation of crystalline nanometer-sized grains within the quasicrystalline matrix [41].

8. Potential Applications

Applications of quasicrystals, both real and anticipated, can be placed into two categories: composites and coatings. (The term composite is used very broadly to include multiple-phase materials produced both by artificial mixing and by natural precipitation.) In both forms, the brittleness which characterizes the bulk material is mitigated. Among the composites, a precipitation-strengthened steel is marketed by Sandvik Steel in Sweden. The precipitates are quasicrystalline. The steel is strong, relatively ductile, corrosion-resistant, and resistant to overaging. The quasicrystalline precipitates are numerous, and resist coarsening during aging. Both of these effects must reflect a low interfacial energy. The steel is attractive for tools in the health industry (surgery, dentistry, acupuncture) and makes up key components of an electric shaver currently marketed by Philips.

Another type of bulk composite is an Al-based alloy which can be formed by rapid solidification and powder processing. The quasicrystals form by precipitation, yielding nanoscale particles surrounded by an Al matrix. Certain compositions can exhibit especially high strength or high ductility [42]. One such alloy, Al – Fe – Cr – Ti, exhibits sufficiently high strength at elevated temperature that it meets an air-force goal level for Al-based alloys. The optimal mechanical properties of these composites, taken over all icosahedra-forming compositions studied to date, are mapped by the outer envelope in Figure 10. The properties are compared with those of a more common Al alloy, for which data are shown by the inner envelope. In all cases (except, of course, thermal expansion), the quasicrystalline-rich composites excel. Gigas, in Japan, markets these materials as extruded bars and preformed parts.

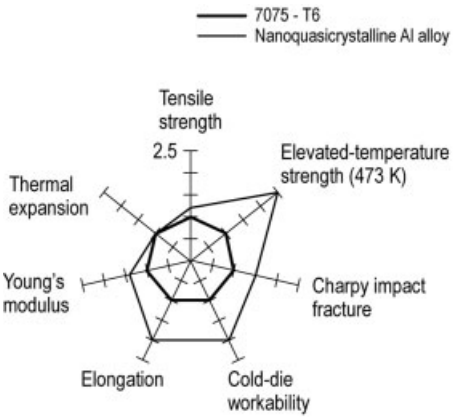

Figure 10. Map of the physical properties of Al quasicrystal composites in comparison with a conventional Al alloy
Courtesy of Akihisa Inoue, Tohoku University, Japan

A type of composite which is promising, but not yet marketed, involves polymers with quasicrystalline fillers. Tests with certain high-performance thermoplastic resins have shown that quasicrystalline AlCuFe particles enhance the wear resistance of the polymers significantly. This effect is not understood, but may be related to a combination of the hardness, low friction, and low thermal conductivity intrinsic to the quasicrystal. At the same time, key thermochemical characteristics of the polymer, that is, its glass and melting transitions, are not degraded, and this indicates that the quasicrystal does not catalyze cross-linking or other disadvantageous reactions in the resin.

Thermal barriers can be manufactured as thick coatings by thermal spraying techniques or magnetron sputtering. This application relies upon bulk properties of quasicrystals, namely, low thermal conductivity and plasticity at high temperature. As such, the quasicrystal is a poor heat conductor which is additionally able to resist the shear stress, if any, generated at the interface with the substrate due to the difference in thermal expansion coefficients. In contrast to zirconia, such a difference is small, since expansion coefficients for quasicrystals, in the range of $(13-16)\times 10^{-6}\,\mathrm{K}^{-1}$, are close to those of metallic alloys. In comparison to metals, quasicrystals also resist oxidation and corrosion by sulfur at high temperature. The main drawback to using quasicrystals comes from their rather low melting temperature and significant atomic mobility. Intercalation of a thin diffusion barrier, typically made of an oxide such as Y_2O_3, allows the latter difficulty to be overcome easily. A specific composition with nearly congruent melting at 1170 °C was developed to enhance the temperature range in which quasicrystalline thermal barriers may be useful. Although not competitive with doped zirconia above 1050 – 1100 °C, such barriers were successfully tested at 950 °C in a real-time ground test of an aircraft engine. Application to other combustion engines, to power generators or, in general, to heat insulation of fast-moving mechanical parts is also of interest.

The potential application for which quasicrystals have gained most attention is as coatings which combine wear resistance with low friction and/or low adhesion. A commercial product which exploited these properties was the quasicrystal-coated frying pan marketed by Sitram. The quasicrystal coating is not directly competitive with Teflon in terms of its nonstick properties, since the work of adhesion of the quasicrystal with water, for instance, is not as small as that of Teflon with water, but adhesion for many foods is reduced relative to a conventional metal pan, and the coating is much harder than Teflon. The hardness means that normal cutlery can be used with the quasicrystal coating, an attractive feature relative to Teflon. Furthermore, the low thermal conductivity also presents an ad-

vantage in the cookware, since it leads to even surface heating. However, it appears that technological problems with the manufacturing process remain to be solved, which points toward the importance of coupling good basic research with technology development in the area of quasicrystals.

The frying pan is a good example of the general need to balance hardness and surface properties in coatings. The place of the quasicrystal in the general scheme is illustrated in Figure 11. It can be seen that the quasicrystal is quite hard, yet offers unique surface properties.

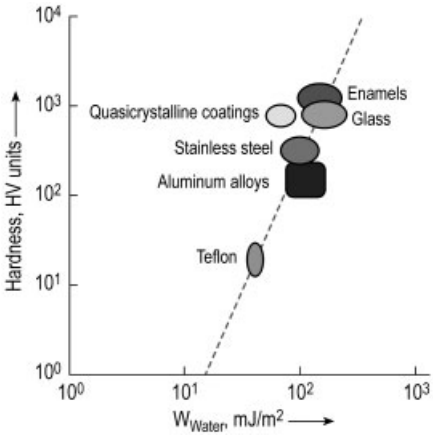

Figure 11. Reversible work of adhesion of water W_{water}, derived from contact angle measurements, as a function of Vickers hardness for various common materials. The size of the symbols shows the range of properties within each family of materials. The straight line shows the relationship expected between hardness and reversible work of adhesion: $H^{1/3}/W =$ constant. Such a selection chart emphasizes the advantage of quasicrystals in terms of striking a balance between adhesion and hardness.

Although one product has been marketed, optimization of coating properties for wear-resistant, low-friction, and/or low-adhesion applications still requires much technological development. One problem is that coatings commonly contain significant levels of phase impurities, which can make them more susceptible to wear by brittle fracture than the pure quasicrystal. The phase impurities arise from the complex solidification pathway of quasicrystals and from other technical problems of fabrication, mostly related to the low thermal conductivity. The wear can also lead to a large increase in the friction coefficient. One possible solution to this problem is the addition of a very small amount of the more ductile phase Fe – Al to a plasma-sprayed coating of Al – Cu – Fe to improve the coating's wear properties. Obviously, the properties of the quasicrystalline coatings are very sensitive to the phase purity, and more specifically to the exact nature of the secondary phase(s).

Some potential applications not mentioned in this article are hydrogen storage, catalysis, thermoelectrics, and optical absorbers [43]. While promising, these applications seem farther from realization than those mentioned above.

Many aspects of the technologies pertaining to quasicrystals have been patented [8, 44–65]. For more scientific details, see [34, 66–69].

In summary, quasicrystals constitute a fascinating scientific frontier, and a barely-explored source of new technologies. A central question, which happens to unite the two frontiers, is this: Why does their remarkable atomic structure lead to unusual physical properties? We believe that the answer to this question will continue to evolve, hand-in-hand with the evolution of new applications.

9. References

We gratefully acknowledge the financial support which enabled this collaboration. In Madrid, this is from Universidad Complutense de Madrid through research program PR64/99-8510. In Nancy, this is from the CNRS, local authorities, and PICS No. 545. In Ames, this is from the National Science Foundation under Grant No. INT-9726785 and through the U.S. Department of Energy under Contract No. W-405-Eng-82. We also thank M. Victoria Hernández and David Hernández for their generous assistance.

Specific References

1. D. Shechtman, I. Blech, D. Gratias, J. W. Cahn, *Phys. Rev. Lett.* **53** (1984) 1951.
2. D. Shechtman, I. Blech, *Metall. Trans. A* **16** (1985) 1005.
3. T. Ishimas, H.-U. Nissen, Y. Fukano, *Phys. Rev. Lett.* **55** (1985) 511.
4. L. A. Bendersky, *Phys. Rev. Lett.* **55** (1985) 1461.
5. N. Wang, H. Chen, K. H. Kuo, *Phys. Rev. Lett.* **59** (1987) 1010.
6. I. Hargittal, *The Chemical Intelligencer* **1997** , 25 – 48.

7. D. Levine, P. J. Steinhardt, *Phys. Rev. Lett.* **53** (1984) 2477.
8. FR 8 810 559, 1988 (J. M. Dubois, P. Weinland).
9. E. Schrödinger, *What Is Life?,* Cambridge University Press, New York 1945.
10. A. Mackay, *Nature* **344** (1990) 21 – 23.
11. C. Janot, M. de Boissieu, *Phys. Rev. Lett.* **72** (1994) 1674 – 1677.
12. C. Janot, *Phys. Rev. B* **53** (1996) 181 – 191.
13. C. Janot, *J. Phys.: Condens. Matter* **9** (1997) 1493 – 1508.
14. C. Janot, J.-M. Dubois in J.-B. Suck, M. Schreiber, P. Hausler (eds.): *An Introduction to Structure, Physical Properties and Application of Quasicrystalline Alloys,* Springer, Berlin 1998.
15. V. Elser, *Phys. Rev. B* **32** (1985) 4892 – 4898.
16. M. Duneau, A. Katz, *Phys. Rev. Lett.* **54** (1985) 2688.
17. P. A. Kalugin, A. Y. Kitaev, L. S. Levitov, *JETP Lett.* **41** (1985) 145.
18. M. Gierer, M. A. Van Hove, A. I. Goldman, Z. Shen, S.-L. Chang, C. J. Jenks, C.-M. Zhang, P. A. Thiel, *Phys. Rev. Lett.* **78** (1997) 467 – 470.
19. C. J. Jenks, P. J. Pinhero, Z. Shen, T. A. Lograsso, D. W. Delaney, T. E. Bloomer, S.-L. Chang, C.-M. Zhang, J. W. Anderegg, A. H. M. Z. Islam, A. I. Goldman, P. A. Thiel in S. Takeuchi, T. Fujiwara (eds.): *Proceedings of the 6th International Conference on Quasicrystals (ICQ6)*, World Scientific, Singapore, 1998, pp. 741 – 748.
20. C. M. H. Beeli, Ph. D. Thesis, Technische Hochschule Zürich (1992).
21. T.-L. Ho in D. P. DiVincenzo, P. Steinhardt (eds.): *Quasicrystals: The State of the Art,* World Scientific, Singapore 1991, pp. 403 – 428.
22. J. L. Aragón, F. Dávila, A. Gómez, *Phys. Rev. B* **51** (1985) 857 – 863.
23. I. R. Fisher, M. J. Kramer, Z. Islam, T. A. Wiener, A. Kracher, A. R. Ross, T. A. Lograsso, A. I. Goldman, P. C. Canfield, *Mat. Sci. Eng. A* **10–16** (2000) 294 – 296.
24. I. R. Fisher, M. J. Kramer, Z. Islam, A. R. Ross, A. Kracher, T. Wiener, M. J. Sailer, A. I. Goldman, P. C. Canfield, *Phil. Mag.* **79** (1999) 425 – 434.
25. I. R. Fisher, K. O. Cheon, A. F. Panchula, P. C. Canfield, M. Chernikov, H. R. Ott, K. Dennis, *Phys. Rev. B* **59** (1999) 308 – 321.
26. A. P. Tsai in Z. M. Stadnik (ed.): *Physical Properties of Quasicrystals,* Springer, Berlin 1999, pp. 5 – 50.
27. W. Hume-Rothery, *J. Inst. Metals* **35** (1926) 295.
28. P. A. Bancel, P. A. Heiney, *Phys. Rev. B* **33** (1986) 7917.
29. J. Friedel, *Helv. Phys. Acta* **61** (1988) 538.
30. A. P. Tsai, A. Inoue, T. Masumoto, *Mater. Trans. J. Inst. Metals* **30** (1990) 463.
31. A. P. Tsai, J. Q. Guo, E. Abe, H. Takahura, T. J. Sato, *Nature* **408** (2000) 537.
32. M. Conrad, F. Krumeich, B. Harbrecht, *Angew. Chem. Int. Ed.* **37** (1998) 1384 – 1386.
33. S. Roche, G. Trambly de Laissardiere, D. Mayou, *J. Math. Phys.* **38** (1997) 1794.
34. Z. M. Stadnik, in M. Cardona, P. Fulde, K. von Klitzing, R. Merlin, H.-J. Queisser, H. Störmer (eds.): *Physical Properties of Quasicrystals,* Springer, Berlin 1999.
35. A. Carlsson, *Nature* **353** (1991) 15 – 6.
36. S. Martin, A. F. Hebard, A. R. Kortan, F. A. Thiel, *Phys. Rev. Lett.* **67** (1991) 719 – 722.
37. E. Belin-Ferré in *Quasicrystals, Materials Research Society Symposium Proceedings* **vol. 553** , Materials Research Society, Warrendale 1999, p. 347.
38. E. Macia, F. Dominguez-Adame, *Electrons, Phonons and Excitons in Low Dimensional Aperiodic Systems,* Editorial Complutense, Madrid 2000.
39. N. F. Mott, *Phil. Mag.* **19** (1969) 835.
40. U. Mizutani, *J. Phys. Condens. Matter* **10** (1998) 4609 – 4623.
41. J. S. Wu, V. Brien, P. Brunet, C. Dong, J. M. Dubois, *Phil. Mag. A* **80** (2000) 1645 – 1655.
42. A. Inoue, H. Kimura, K. Kita in A. I. Goldman, D. J. Sordelet, P. A. Thiel, J. M. Dubois (eds.): *New Horizons in Quasicrystals: Research and Applications,* World Scientific, Singapore 1997, pp. 256 – 263.
43. P. C. Gibbons, K. F. Kelton in Z. M. Stadnik (ed.): *Physical Properties of Quasicrystals,* Springer, Berlin 1999, pp. 403 – 432.
44. FR 9 100 549, 1991 (J. M. Dubois, A. Pianelli).
45. FR 9 115 866, 1991 (J. M. Dubois, P. Archaumbault, B. Colleret).
46. FR 9 215 659, 1992 (J. M. Dubois, P. Carthonnet).
47. SE 9 303 280, 1993 (A. H. Stigenberg, J. O. Nilsson, P. Liu).
48. DE 4 425 140, 1994 (T. Eisenhammer, M. Lazarov).
49. FR 9 503 938, 1995 (F. Cyrot-Lackmann, T. Grenet, C. Berger, G. Fourcaudot, C. Gignoux).

50. FR 9 503 939, 1995 (J. M. Dubois, F. Machizaud).
51. JP 24 325 392, 1992 (K. Kita).
52. JP 20 662 692, 1992 (A. Inoue, T. Masumoto, J. Nagahora).
53. JP 6 256 692, 1992 (T. Masumoto, A. Inoue, M. Wattanabe, J. Nagahora, T. Shibata).
54. JP 26 559 193, 1993 (K. Nosaki, T. Masumoto, A. Inoue, T. Yamaguchi).
55. JP 32 437 393, 1993 (T. Sawa, T. Inaba, Y. Takahashi).
56. Samsung Electronics, Ltd., US 5 914 594, 1994 (T. Masumoto, A. Inoue, H. Kimura, Y. Shinohara, Y. Horio, K. Kita).
57. Sandvik, AB 5 632 826, 1997 (A. Hultin-Stigenberg, J.-O. Nilsson, P. Liu).
58. Iowa State University Research Foundation, US 5 851 317, 1998 (S. B. Biner, D. J. Sordelet, B. K. Lograsso, I. E. Anderson).
59. Centre National de la Recherche Scientifique, US 5 204 191, 1993 (J.-M. Dubois, P. Weinland).
60. US 5 397 490, 1995 (T. Masumoto, A. Inoue, J. Nagahora).
61. YKK Corp., US 5 419 789, 1995 (K. Kita).
62. Centre National de la Recherche Scientifique, US 5 432 011, 1995 (J.-M. Dubois, A. Pianelli).
63. Iowa State University Research Foundation, US 5 433 978, 1995 (J. E. Shield, A. I. Goldman, I. E. Anderson, T. W. Ellis, R. W. McCallum, D. J. Sordelet).
64. US 5 458 700, 1995 (T. Masumoto, A. Inoue, M. Watanabe, J. Nagahora, T. Shitata).
65. YKK Corp., US 5 607 523, 1997 (T. Masumoto, A. Inoue, J. Nagahora, T. Shibata, K. Kita).
66. A. I. Goldman, D. J. Sordelet, P. A. Thiel, J. M. Dubois (eds.): *New Horizons in Quasicrystals: Research and Applications,* World Scientific, Singapore 1997.
67. D. J. Sordelet, J. M. Dubois, *MRS Bulletin: Quasicrystals,* **22** (1997).
68. C. Janot, J.-M. Dubois, *Les Quasicristaux. Matière à Paradoxes,* EDP Sciences, Les Ulis 1998.
69. J.-M. Dubois, P. A. Thiel, A.-P. Tsai, K. Urban (eds.): *Quasicrystals, Materials Research Society Symposium Proceedings* **vol. 553** , Materials Research Society, Warrendale 1999.

Quinacridone Pigments → Pigments, Organic

Techniques for DNA Analysis

WILSON J. WALL, Kidderminster, UK

1. Introduction 1
1.1. DNA Structure 1
1.2. Structure of DNA in Vivo 3
2. Primary Molecular Tools for DNA Analysis 3
3. Methods of DNA Detection 6
3.1. Bioluminescence 7
3.2. Colorimetry 7
3.3. Electrochemiluminescence 7
3.4. Fluorescence 7
3.5. Denaturing Gradient Gel Electrophoresis (DGGE) 10
3.6. Single-Strand Conformation Polymorhism 11
3.7. Random Amplified Polymorphic DNA 11
3.8. Short Tandem Repeat Analysis . . . 11
3.9. Single Nucleotide Polymorphism (SNP) Detection 12
3.10. Mitochondrial DNA Analysis 13
3.11. DNA Analysis and Bioinformatics . 14
4. Applications of DNA Analysis 14
4.1. General Principles of ARMS Analysis of DNA 15
4.2. Analysis of Dynamic Mutations . . . 16
4.3. Using DNA Analysis to Determine Sex 17
4.4. Methods of Personal Identification . 18
4.5. Bacterial Contamination of Water Supplies 19
4.6. Adulteration of Food Stuffs 19
5. References 20

1. Introduction

DNA is a biological polymer made from very simple building blocks but exhibiting large-scale complexity. It is the complexity which lends this remarkable molecule its power to control cells, tissues, and ultimately organisms. It is also this complexity which is looked at when DNA is analyzed. The length of the molecule gives a sense of scale. For example, each human chromosome contains DNA varying between 1.4 and 7.3 cm in length, depending on the chromosome.

1.1. DNA Structure

DNA is composed of a chain of nucleotides, each of which is made up of a deoxyribose sugar, a phosphate group, and a base. Variety in DNA stems from the bases; these can be either adenine or guanine (purines), or cytosine or thymidine (pyrimidines). The phosphates and sugars provide the external backbone from which the bases project. The backbone is constructed of sugars attached to each other by formation of a phosphate ester between the 3′-hydroxyl group of one sugar and the 5′-phosphate group on the next sugar; the DNA molecule is therefore polarized 5′ to 3′.

The double-helix structure is formed by two strands of DNA pairing in opposite polarities. The phosphate – sugar backbone is on the outside, and the bases project into the middle. While the components of the individual strands are held together by covalent bonds, the two strands are held together by hydrogen bonds between the bases.

The four bases associated with DNA, which are generally referred to by their initial letters, are adenine (A), cytosine (C), thymidine (T), and guanine (G). These bases key together from opposite strands in a very particular way. A only pairs with T, and G only pairs with C. The association is controlled by formation of hydrogen bonds (Figure 1).

It is the delicacy of the hydrogen bonds compared with covalent bonds which lends both functional performance and elegance to DNA. The conformation of the double stranded helix can be visualized as a right-handed spiral staircase with 10 steps (bases) per complete revolution and a diameter of approximately 2 nm. As a consequence of this structure, it is not possible to separate the two strands without unwinding the helix.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.e26_e01

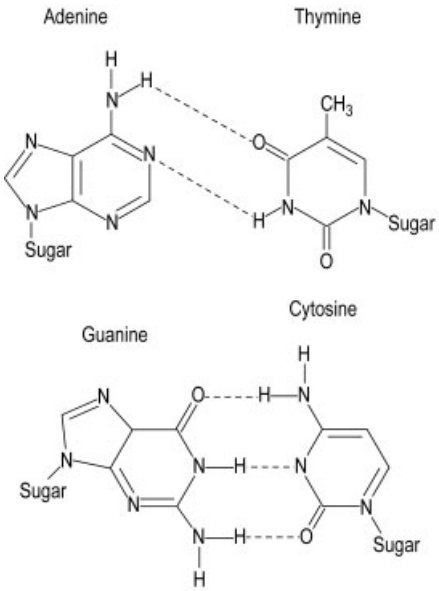

Figure 1. Normal base pairing found in DNA between adenine and thymine and between cytosine and guanine

It is a useful convention when dealing with DNA to ignore the sugars and phosphates and merely refer to the bases of a single strand. This is because the exact bonding of one strand with another allows the complementary sequence of any given single strand to be precisely determined. Consequently, when a DNA sequence is written out, it is conventional only to list the base sequence of a single strand, the other being explicitly determined. Note that genetic complexity is not directly related to DNA content of a cell. This can be broadly seen in Table 1.

Table 1. DNA content in picograms of a range of species

Species	Common name	Nuclear DNA content, pg
Saccharomyces cerevisiae	yeast	0.026
Drosophila	fruit fly	0.1
Mus musculus	mouse	2.5
Homo sapiens	man	3.7
Avena sativa	bread wheat	18.1
Protopterus	lungfish	50
Fritillaria davisii	lily	98.4

The huge differences in DNA content are generally associated with variations in the amount of repeated sequences present. The complexity of the genome, that is, the proportion of repeats compared with a unique sequence in any given organism, is determined by using Cot curves, Cot being the concentration (Co) of DNA in moles of nucleotide per liter multiplied by the renaturation time (t) in seconds. Calculation of Cot values is complex, but they are always presented in a uniform way. The values are plotted on a log scale to give a curved plot whose slope gives an indication of the level of repetition in the DNA sample. Cot curves are based on the fact that DNA can be denatured by heat and reanneals when the temperature drops. This behavior also forms an important part of many other methods of DNA analysis.

When DNA is cut into short sections of about 400 – 600 bp (base pairs) and these fragments are then denatured, the highly repetitious sequences reanneal first, simply because the repeat sequences are more likely to come into contact with complementary sequences. The more complex the genome is the longer it takes low- or single-copy sequences to find the right partner.

1.2. Structure of DNA in Vivo

DNA does not sit in cells in an uncontrolled manner. It is regulated in both its position and expression. These are carried out by proteins, genetically silent, but of huge importance. The proteins involved in structural control of DNA are some of the most highly conserved found in the living world. While it is not the remit of this article to detail the manner in which DNA is controlled within the cell, it is worth noting the first-order structure of chromatin, this being the term used to describe the DNA – protein complex.

By using osmotic shock it is possible to unwind chromatin to show the fundamental unit as a beaded string of about 10 nm in diameter. The beads are histone proteins, around which the DNA double helix is wrapped. Histone proteins are almost exclusively associated with nuclear DNA. The five different histones associated with DNA, designated H1, H2A, H2B, H3, and H4, have been highly conserved throughout evolution, and this indicates their fundamental importance to the integrity and control of DNA. Changes to histone proteins by methylation, phosphorylation, etc., have effects on the charge of the protein, which can alter the interaction between histone and DNA. For example,

acetylation of histone H4 results in unfolding of the nucleosome core and is associated with transcriptionally active regions of DNA.

Beyond this level of organization the DNA – histone complex is repeatedly coiled such that the whole strand becomes shorter and shorter and more tightly controlled. The ultimate winding step results in a chromosome, which during certain stages of the cell cycle is sufficiently well condensed to be visible with a light microscope, and specific staining techniques can be used to identify individual chromosomes. Visualizing chromosomes is relatively easy in eukaryotes, that is, higher organisms with a clearly defined nucleus. However, in prokaryotes such as bacteria, the nucleic acid is not generally associated with histone proteins, and single chromosomes are often circular. This reflects the simplicity of the genome. For example, humans, who do not have the largest genome in the animal kingdom, contain approximately 3000 million base pairs, while the gut bacterium *Escherichia coli* contains 4 720 000 base pairs and a virus which infects *E. coli* with the designation ΦX174 contains 5386 bases of single-stranded DNA.

2. Primary Molecular Tools for DNA Analysis

With DNA being such a large molecule it is important to be able to manipulate it effectively. This can take any form of manipulation from the entire chromosome, or to just very small pieces of DNA of particular interest. To make these manipulations various enzymes are used which are a part of a cell's own ability to alter the activity of DNA and in some cases protect the cell from invasion by pathogens.

The most important types of enzyme used in molecular biology are:

- Exonucleases
- Endonucleases
- Polymerases
- Ligases
- Methylases

Exonucleases and endonucleases are collectively referred to as DNases, or more fully deoxyribonucleases. Enzymes which can degrade both DNA and RNA are called simply nucleases. If a DNase cuts only one strand of a double-stranded molecule it is called a nicking enzyme, and the resultant single-strand break is a nick. One such enzyme is Dnase I. These enzymes seem to be important in releasing tension in the double helix as it is unwound during replication.

Exonucleases degrade DNA from the ends and cannot themselves generally introduce breaks within a strand. For example, mung bean nuclease progressively remove bases from both the 5′ and 3′ ends of DNA where there is an overhang, resulting in two blunt ends, that is, strands of equal length of complementary sequences. In contrast, exonuclease III preferentially degrades recessed 3′ ends, that is, where the complementary strand extends beyond the 3′ end. If the 3′ end overhangs by more than four bases, the enzyme cannot attack it. This property can be exploited by first nicking the DNA at specific sites with an enzyme and then using exonuclease III to create a single-strand gap propagating in a specific direction.

Endonucleases cut DNA internally, rather than at the ends, and are more correctly called restriction endonucleases. The activity of endonucleases can vary widely from enzyme to enzyme. Several hundred are now known, with very different operating conditions and sites at which they restrict, or cut, the DNA. Endonucleases recognize short sequences along the DNA, which may occur once every few hundred or few thousand bases along the molecule. The frequency with which cutting sites can be found depends largely upon either how specific the sequence must be or how long the sequence must be before it is recognized by the enzyme.

Most endonucleases recognize sites between four and six bases long and cut the two strands either symmetrically or staggered. A range of such restriction endonucleases and their restriction sites are shown in Table 2.

Table 2. A range of enzymes and their cutting sites

Enzyme name	Restriction site
Nae I	5′ . . .GCC↓GGC. . .3′
	3′ . . .CGG↑CCG. . .5′
Nco I	5′ . . .C↓CATGG. . .3′
	3′ . . .GGTAC↑C. . .5′
Hae III	5′ . . .GG↓CC. . .3′
	3′ . . .CC↑GG. . .5′
Mbo I	5′ . . . ↓GATC. . .3′
	3′ . . . . . .CTAG↑ . . .5′
Msp I	5′ . . .C↓CGG. . .3′
	3′ . . .GGC↑C. . .5′

The names are generally derived from the initials of the organism in which the particular enzyme was found; for example, all restriction enzymes beginning *Eco* are derived from *Escherichia coli.*

Using any given endonuclease on a specific piece of DNA always gives the same number and size of fragments. With large genomes there can be variation between individuals, but with very small genomes, as found in viruses where there is no sexual reproduction, the variation in restriction sites between individuals is effectively zero. So using a restriction endonuclease on such material results in fragments of exactly known size. These could then be compared with unknown fragments to give information about the DNA under analysis. An example of this is the digestion of a small plasmid of 4361 base pairs, designated pBR322, which is usually found in *E. coli*. Digestion of this with the endonuclease *Msp* I always results in 26 fragments ranging in size from 622 base pairs (4.04×10^5 Da) to only nine base pairs (0.06×10^5 Da).

Polymerases take individual nucleotides and attach them to a progressively lengthening chain of DNA. In living cells they perform the essential function of DNA replication. This is also exploited in vitro to produce multiple copies of specific regions of DNA. In cells, polymerase is controlled by a battery of associated enzymes. All known polymerases replicate in the $5' \Rightarrow 3'$ direction. In the laboratory, a different approach must be taken to make full use of the technique.

There are many different types of DNA polymerase, most of which are, like most enzymes, unstable at high temperatures. These enzymes are of a specific value in some circumstances. For example, T7 DNA polymerase comes from bacteriophage T7, which attacks *E. coli* by taking over cellular control and replicating its own DNA. This enzyme can be very useful for replicating long stretches of DNA quickly. Sometimes a short section of DNA must be replicated, but more than once, and for this thermostable enzymes are needed. The first of these was extracted from a deep sea species *Thermus aquaticus* and called *Taq*. Later other forms of heat-stable DNA polymerases were found in other species living around deep-sea thermal vents. These enzymes tend to make fewer errors in replication. They include a DNA polymerase from *Thermococcus litoralis*, which lives in a marine environment at 98 °C, and one extracted from a species of *Pyrococcus*, recovered from a deep-sea vent at 2010 m and capable of growth at 104 °C.

The value of these thermostable enzymes is that they can easily be used in polymerase chain reactions (PCRs), a technique in which short sections of DNA are replicated exponentially (see also → Polymerase Chain Reaction). By knowing two short sequences of DNA on either side of the target DNA, which itself can be of a completely unknown sequence, it is possible to produce single-stranded primers, complementary to the two sequences. At this stage the target DNA is double stranded, held together by hydrogen bonds. Exposure to heat denatures the DNA and results in single-stranded DNA. On cooling the sample at this stage, to a temperature at which the DNA will reanneal, some complementary strands will rejoin, but many of the primers will attach to the single-stranded DNA. Increasing the temperature at this point to below the denaturing temperature allows the DNA polymerase to extend the strand by adding on bases that are also present in the solution. This cycle is repeated until enough target DNA is present for further analysis [1].

PCR is one of the most powerful methods of producing material for analysis from very small samples. If PCR were 100 % efficient, then after 20 cycles there would be 2^{20} times the number of starting molecules, each with a primer at one end. It is these primers which can be tagged, if required, for later detection. Every primer detected which exceeds the original primer in mass implies the presence of an amplified product. After approximately 25 – 30 cycles the enzyme

limits the reaction and so, in conjunction with other factors, this results in a realistic 10^6-fold increase in target DNA. One such limiting factor is that as PCR proceeds and more complementary DNA is produced, so there is an increasing tendency for complementary strands to anneal, to the exclusion of the primers. Figure 2 illustrates the temperature cycles and the increase in amplified product.

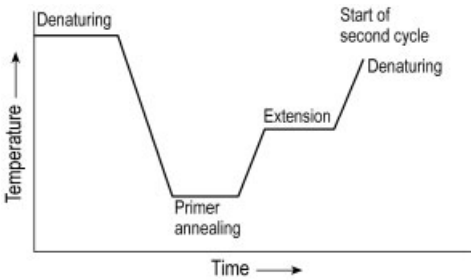

Figure 2. The polymerase chain reaction temperature cycle. The exact temperatures depend upon the type of primers; AT-rich primers are more easily denatured than GC-rich primers.

Ligases reattach strands of DNA together. Ligases come in many forms, some only able to work on single-stranded DNA or RNA, while others act on duplex strands. Another major difference between ligases is the ability to reseal a single-strand nick created by a nicking enzyme. Although some ligases only reattach blunt ends, single-base overhangs, or cohesive ends, that is, two single-strand overhangs with exactly complementary sequences, all ligases operate in the same way. The enzyme catalyzes the formation of a phosphodiester bond between juxtaposed 5′-phosphate and 3′-hydroxyl termini of two adjacent oligonucleotides.

Methylases chemically modify DNA residues. In higher eukaryotes this usually takes the form of attachment of a methyl group to cytosine to form 5-methylcytosine. Usually the methylated residues are in C – G pairs. The level of methylation of eukaryote DNA shows a strong correlation with transcriptional inactivity. Lower eukaryotes and prokaryotes are also able to methylate adenine residues. There are no known methylases that act on guanine.

Methylases have recognition sites in which they work, so it is generally possible to tailor the methylation to a specific area of DNA. The value of this stems from methylated residues being resistant to attack from restriction endonucleases. This allows the bacteria carrying endonucleases for protection against invading viruses to protect their own DNA by methylation. It is this aspect of methylases which is important in DNA analysis.

3. Methods of DNA Detection

Once a DNA sequence of interest has been produced by PCR or extracted from a cell, there are several further stages which can be used to determine more about the DNA. Broadly speaking, two aspects can be investigated: size and sequence. Often, for practical purposes, sizing alone is sufficient to detect, say, a mutation of medical importance [2]. In some cases, sizing of some specific fragments may be enough to allow determination of a sequence.

Sequencing is generally undertaken only when absolutely necessary, because of cost and time constraints. The best known sequencing effort is the Human Genome Mapping Project (HGMP). This is intended to sequence the entire human genome, but it should not be imagined that the story ends there. The reading frame of much of the DNA is still unknown, and it is also unknown what most of the sequences are for, if anything. Consequently, knowing a sequence is only the start.

There are several different ways in which DNA can be detected and sized, the majority of which involve separation of different sized DNA fragments under the influence of an electric current (electrophoresis). The medium for gel electrophoresis varies from simple agarose to polyacrylamide.

The earliest method for detecting DNA fragments after electrophoresis was the Southern blot. It involves transfer of the various discrete fragments to a solid support, where they are immobilized. This transfer is accomplished by capillary action in SSC (sodium chloride – sodium citrate) buffer. The capillary action transfers the DNA to a nylon or other artificial membrane, after which it is fixed by heating to approximately 80 °C. The transfer takes place on an horizontal

bed, so the relative positions of the DNA does not alter. Treating the DNA on the blot with probes that have complimentary sequences to those which are under investigation yields a series of bands that can be taken as representing the positions of those known, probe, sequences of DNA. These probes are not irreversibly bound to the membrane, unlike the DNA under investigation, so they can be removed, and a different probe, or probes, hybridized.

Broadly, there are two different types of probe: multilocus probes (MLP) and single-locus probes (SLP). Multilocus probes hybridize at several different sites on the genome, while single-locus probes only hybridize to a single site. Hence, an MLP will give a multi-banded result, but an SLP will give a maximum of only two bands per probe per individual.

The reason for the maximum of two bands is that all higher organisms carry two copies of every gene, one from each parent. When using SLPs these may be both the same (homozygous) and result in only a single band. However, when the two alleles are different (heterozygous) an SLP will produce two bands. Whichever of these systems, SLP or MLP, is used, indeed, whenever a gel separation system is used, an internal size standard is required. This either an allelic ladder or DNA digested by a specific enzyme yielding fragments of known sizes. The comparison of fragment sizes is generally referred to as the Local Southern method. It takes into account the tendency of DNA fragments that are similar in mass to migrate in a generally linear fashion relative to each other. It therefore requires at least three DNA mass markers which are close in size to the fragment of interest so that an accurate comparison can be made. It is important that the DNA mass markers are closely associated with the fragments under investigation because as they depart in mass, the calculation of mass by comparison becomes increasingly inaccurate.

This technique of comparison of size of DNA fragments against a standard is still the best and most accurate method of determining the size of a piece of DNA when it is either fixed on a membrane or travelling through a gel. This extends to other techniques which do not require the use of a membrane blot to transfer DNA to a solid substrate.

While the original method of DNA analysis required the transfer to a membrane of DNA that had been separated by gel electrophoresis before further analysis could be undertaken, more recent developments have rendered this unnecessary except in specific areas of research. However, automated systems still require a method of sizing which is both reliable and repeatable, and so DNA fragments of known mass are still the method of choice.

Techniques for detection of DNA on gels has also changed. In the original method, ^{32}P, which has a half-life of approximately 14 d, was substituted for a nonradioactive P atom in the DNA probe. When hybridized to the probe and exposed to X-ray film, the film would darken at the radioactive sites. Besides the short half-life, two other major considerations spawned research into nonisotopic detection of nucleic acids: potential long-term health hazards and disposal of residual radioactive material from the assay. Most chemical detection techniques were based around immunoassay tests and then developed further.

Nonisotopic labeling to detect nucleic acid – probe hybrids can direct or indirect [3]. Direct methods involve attaching a detectable label directly to the target, usually the probe, by a covalent bond. This technique is now employed in machine-based systems, but on gels it is still common to use an indirect method. This involves a hapten, a low molecular mass molecule (usually < 1000), attached to the probe, which can then be detected by using a specific binding protein, which may be antibody based. Although the binding strategy should avoid involving the hydrogen bonds, which are essential for binding of probes to the target DNA, this is not always practicable and may partially compromise the efficacy of the system.

The sensitivity of hybridization assays is a function of the detection limit of the label, some of which are not only as sensitive as radiolabeled assays, but are much faster in exposure times to X-ray film. There are four broad categories of detection for nonisotopic labels: bioluminescence, colorimetry, electrochemiluminescence, and fluorescence.

3.1. Bioluminescence

This is essentially the same as chemiluminescence found in nature, for example in fireflies

and certain types of fungi and bacteria. They involve the use of a luciferin substrate and an enzyme, luciferase. Because of the versatility and sensitivity of these systems, they can be used with either film or charge coupled device (CCD) cameras.

3.2. Colorimetry

Systems using colorimetry tend to be less sensitive than luminescent systems, but they do have some advantages. By producing a precipitate associated with the DNA – probe hybrid, they are ideal to produce a permanent record of the distribution on a membrane.

3.3. Electrochemiluminescence

Electrochemiluminescence is a specialized technique in which an electrochemical reaction produces an excited state, after which decay to a ground state produces light. The special facilities for electrochemical generation and detection renders this type of system expensive.

3.4. Fluorescence

Fluorescence detection is the system of choice now generally employed in automated equipment. Straightforward fluorescence detection of DNA on gels by film or CCD device, is difficult because of intrinsic problems of noise. However, specific fluorochromes have enabled systems to be developed which utilize different methods of detection, rather than simply overlaying a gel with a photoreactive paper.

Fluorescence detection of specific dyes at specific wavelengths has revolutionized not only the detection of target alleles, but also personal identification, paternity details, and DNA sequencing. This technology has been implemented in many different systems using a range of equipment. A common method for detection of specific fragments is to attach a dye to either an oligonucleotide primer, for use in a PCR reaction, or a probe. By using different fluorochromes for different oligonucleotide primers it is possible to run several PCR reactions simultaneously in the same reaction vessel. Detection involves inducing fluorescence with a laser, the data being captured and processed automatically as the DNA fragments pass in front of the laser. Thus, an image can be created from the data which broadly corresponds to the older autoradiograph image created from Southern blots and probes. However, this is not actually the most convenient method of data display. The graphic display in Figure 3 is the more usual method of showing results of analysis.

By using four different dyes (red, blue, green, and yellow), large numbers of reactions can be carried out in a single tube. It is also possible to include size standards, which makes accurate sizing of fragments much more precise. Figure 4 shows the emission spectra of the various dyes in commercial use. Even if two different fragments have the same molecular mass, povided they fluoresce at different wavelengths they can be separated by using filters. Since many of these dyes are proprietary products, their exact chemical constitution has not been released.

So far these methods have broadly been applicable to slab gels, but more recently capillary electrophoresis for DNA has been introduced. The main advantage of this system is speed and reproducibility. Arrays of capillaries are used instead of a single gel, but in general terms the technology is very similar and the output identical.

As new technologies are introduced there is a move away from gel-based systems towards more exotic methods of DNA analysis. One such is the DNA chip. This is a technology still in its infancy, but potentially of great power and speed. DNA chips utilize the negative charge on DNA to move it about and concentrate it in target areas. These target areas have probes chemically bound to the chip. After several cycles of DNA binding to the chip, each with a different probe, the same process is repeated, but this time with the DNA to be analyzed. By concentrating the DNA around the probes, hybridization rapidly takes place. The bound DNA is detected with a laser and arrays to detect the resulting fluoresence. As can be readily appreciated, although potentially very useful, the difficulty is in making the chips flexible enough to be applicable to any but the most specialist of areas of research and diagnostics, where mass screening would make the system commercially viable.

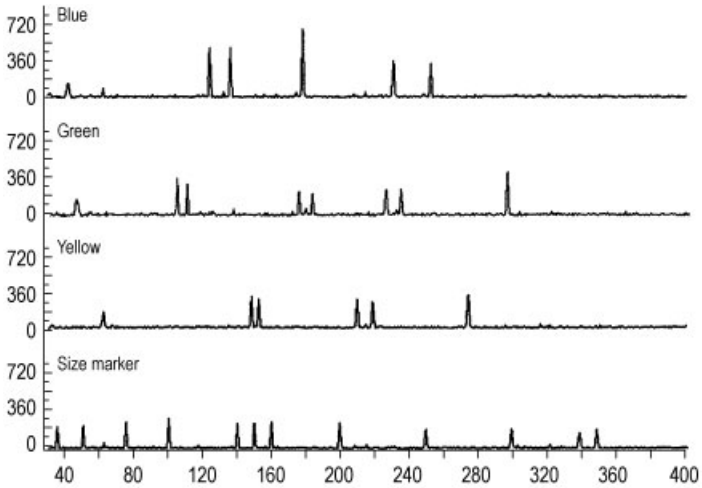

Figure 3. Graphic display of output from an automated DNA analyzer. By using loci of widely differing sizes in the same lane, they can use the same laser-activated dye without risk of confusion. This is a composite image of data generated on the basis of when each fragment passes the laser window. It is therefore not a static image but based on a dynamic timescale. The higher molecular mass fragments are towards the right-hand side of the display. Note a size maker is still needed, at the bottom; itfluoresces red. Since all the alleles are separable, they can be run in a single column, the image being generated by the associated computer.
Image courtesy of PE Applied Biosystems

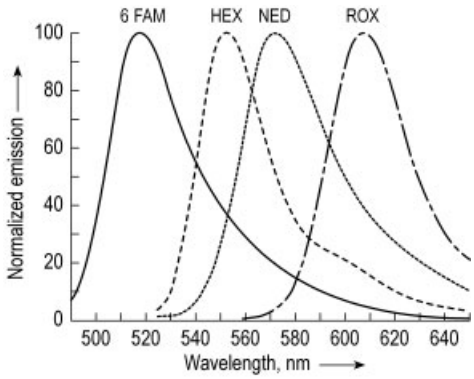

Figure 4. The emission spectra of four fluorescent dyes commonly used in automated DNA analyzers.
Image courtesy of PE Applied Biosystems

One of the most useful techniques to help in understanding DNA and genes is DNA sequencing, which determines the precise arrangement of bases making up the DNA. This has received considerable public exposure with the advent of the Human Genome Mapping Project (HGMP), an attempt to sequence the entire human genome. Although the implementation of new fluorescence-based technology has speeded up and automated the process of DNA sequencing, the method used is still broadly based on the Sanger–Coulson method of chain-terminating nucleotides. It is therefore the detecting methods that have changed, rather than the underlying science. The alternative method of sequencing, chain degradation, has not been so widely employed in modern systems.

The chain-terminating method of DNA sequencing requires single-stranded DNA, because it involves the production of a second complementary strand. Consequently, the DNA to be sequenced is normally cloned into a single-stranded vector such as a plasmid. This is done by cutting out the gene of interest by using nucleases and then ligating it back into the plasmid. By having single-stranded DNA, both to be sequenced and in the vector, it is possible to produce a short oligonucleotide primer which will anneal to the plasmid and allow extension of the DNA by using the gene under investigation as the template for the growing PCR product. The use of a short primer is necessary because most of the DNA polymerases which are used in sequencing reactions need a double-stranded region from which to start synthesis.

A chain-terminating sequencing reaction requires each of the four deoxynucleotides which

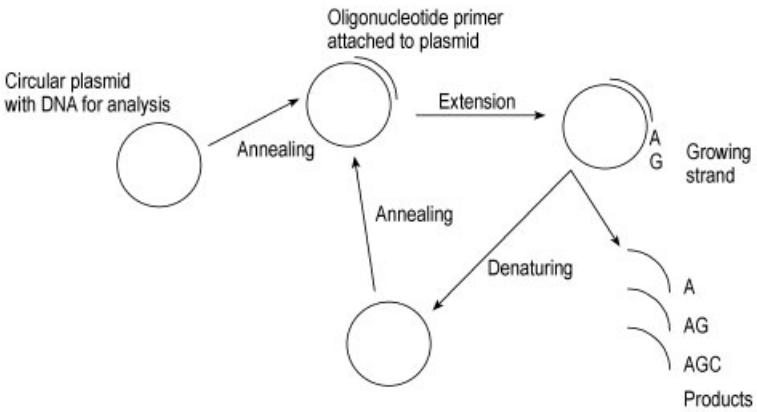

Figure 5. The circular plasmid, containing the DNA to be sequenced, short oligonucleotide primers complementary to a known section of DNA, and labeled nucleotides are reacted together. The primer attaches to the plasmid, and then polymerase enzymes extend the primer until incorporation of a dideoxynucleotide terminates chain extension. Denaturing the products, followed by annealing of more primer results in accumulation of artificially produced single strands of DNA complementary to the original template. Overall there will be DNA strands varying throughout the total area of interest in steps of only one base, allowing for step by step sequencing.

can be incorporated into the growing strand. These are dATP, dTTP, dGTP, and dCTP. At this point the original method would require four reactions to be carried out separately. One with the addition of dideoxyATP (ddATP), one with ddTTP, one with ddGTP, and one with ddCTP. These are the chain terminators. Dideoxynucleotides can be incorporated into a growing chain as easily as a deoxynucleotide, but block further strand synthesis. This blockage happens because the dideoxynucleotides lack the hydroxyl group at the 3′-position of the sugar, where the next nucleotide would normally be attached.

When ddCTP is added to the mixture, strand growth will cease when it is incorporated opposite a G on the template strand. However, by adjusting the concentrations of the reagents in the reaction mixture, termination will not always take place at the first template G, because an ordinary dCTP may be added by the enzyme instead. So a number of different fragments of different lengths are produced from this one reaction, each terminated by a ddCTP. When it was normal practice to radiolabel the dideoxynucleotides, they had to be run in separate reactions and run in adjacent columns on a sequencing gel, which was then blotted, autoradiographed, and the bands read horizontally across the gel lanes, with only one band in any position down the gel representing the base at that position on the strand. Sequencing gels are very thin polyacrylamide containing urea to denature the strands.

The modern trend is towards labeling the dideoxynucleotides with fluorescent dyes, each fluorescing at a different wavelength [3]. Besides allowing the sequence reading part of the experiment to be carried out automatically, it also means that the four different reactions can be carried out in one tube, and sequence reading carried out in one lane of a gel [4]. This automation of processes helps to minimize errors. Since the reaction is competitive between deoxynucleotides and dideoxynucleotides, there is a tendency for more low molecular mass products to be formed than high molecular mass ones. There is also a limit to the length of strand which can be accurately sequenced, because as they get longer the percentage molecular mass difference between strands differing in length by a single base becomes progressively smaller. Figure 5 shows diagramatically how cycle sequencing is carried out with a circular plasmid and a thermostable DNA polymerase, so that the reaction can be set up and left to run without intervention until the results are transferred to an electrophoresis system.

Figure 6 shows four lanes of a sequencing gel. As the molecular mass of the fragments increases, towards the top of the gel, they move more slowly and become increasingly difficult

to separate. Using fluorescent dyes allows all the products to be run in the same gel lane. Although the DNA fragments still tend to "bunch" as their molecular mass increases, it is easier to separate the different products because of their different fluorescent wavelengths.

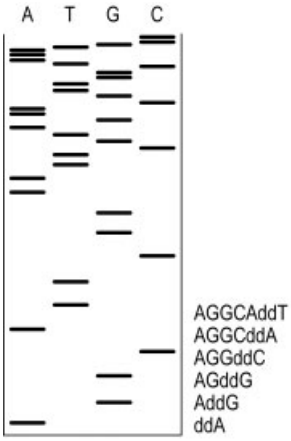

Figure 6. A sequencing gel produced from four separate reactions in a chain termination experiment. Note how these are read sequentially across the gel, and as the molecular mass increases band separation becomes more difficult. By using different colored fluorescent dyes, these reactions can be carried out in the same tube and sequenced on the same track of a gel, or simultaneously on a single capillary electrophoresis device

3.5. Denaturing Gradient Gel Electrophoresis (DGGE)

It is possible to determine the presence or absence of a DNA polymorphism without actually knowing anything about the nature of the polymorphism itself. This is done by using DNA which has been through a PCR to produce enough material to be analyzed. The detection relies on the fact that DNA heteroduplexes, which differ from each other by only a single base pair, not of content, but of type, have measurably differing characteristics that affect the transition from double-to single-stranded DNA. Since single-stranded DNA travels at a slower rate than double-stranded DNA during electrophoresis through polyacrylamide gels, a denaturing gradient will enhance separation.

By having a gradient of a DNA-denaturing agent through the gel, as the similar, but not identical, double-stranded DNA passes down the gel it will melt at different points, at which point the passage through the gel slows down and the original heteroduplexes can be seen to be different. Of course, it is not possible to say in what way the heteroduplexes are different, only that they are. The denaturing gradient can be introduced in a number of ways, such as formamide, urea, or a temperature gradient.

3.6. Single-Strand Conformation Polymorhism

Similar to DGGE, single-strand conformation polymorphism (SSCP) is a method which can be used to detect mutations and polymorphisms without characterizing them. SSCP analysis utilizes the principle that electrophoretic mobility is dependent on size and shape of the molecule and the electric charge. Ideally DNA to be analyzed by SSCP has less than about 400 base pairs. By running a PCR across the area to be tested, enough DNA can be produced for analysis. This is carried out by denaturing the PCR products and then rapidly cooling them. This stops most of the single-stranded DNA reassociating with complementary sequences, but does allow the single-strands to fold into a conformation determined by the sequence itself. A difference of only a single base will alter the conformation and therefore the mobility within the gel, which is of a native, rather than a denaturing, type. Because the conformation can be affected by ion concentration and temperature it is important that the system is highly uniform throughout the proceedure. This means controlling the temperature of the gel during the experiment as well as the solvent and ion concentrations.

3.7. Random Amplified Polymorphic DNA

Random amplified polymorphic DNA (RAPD) is in some ways a misleading name. It allows detection of polymorphisms without any knowledge of the polymorphism, but also without any knowledge of the DNA sequence of the target organism. Indeed, the results can be of limited value because of this, but there are specific applications to which this technique can be put with

great success. RAPD analysis requires the use of a single short oligonucleotide, usually about ten bases long, the sequence of which is chosen at random. The primer will often anneal to genomic DNA during PCR at several different places, giving rise to several bands on subsequent electrophoresis, but if conditions are maintained between experiments the number of bands should remain the same for any individual.

The amplified fragments are highly dependent on the primer sequence and the origin of genomic DNA being looked at. So the results may be of interest, but without further analysis it is not possible to say what form the polymorphisms take or where in the genome they reside. One area of particular interest for RAPD use is in agriculture where it might be necessary to determine the sex of a plant in dioecious species, that is, species for which the male and female flowers occur on different plants. This is useful to stop the planting of more pollinating plants than is necessary, but, as is the case with date palms, until the trees mature and flower it is not possible to determine which trees are male and which female. So by using RAPD analysis differences may come to light between the sexes which can be exploited to determine the sex of the plant while still a seedling.

3.8. Short Tandem Repeat Analysis

Short tandem repeats (STR) are widely distributed throughout the genomes of virtually every higher organism that has been investigated. STRs are arrays of variable numbers of short repeated sequences. They are sometimes referred to as microsatellites since they are very small compared with other satellite DNA. Satellite DNA can be defined as DNA from a eukaryote that separates as a distinct fraction during gradient centrifugation in cesium chloride. The satellite fraction is made up of material of a distinct composition, having a preponderance of A+T or G+C. Some satellite DNA, such as the Alu sequence found in humans, is made up of 300 bp repeats, but in STRs the repeat may be only 1 – 6 bases long. The repeat can be made up of any combination of bases, repeated tandemly.

STRs can be found scattered through the genome in most, if not all, genes. They are located in intron DNA, that is, the interspersed noncoding sections which lie between the exons, or coding sequences. While it is true that introns are not transcribed, it would be incorrect to call them junk DNA. Although not as yet clear why, it seems that introns are necessary to maintain the integrity of the gene. They are also important in evolution, allowing changes to take place within the genome without compromising the metabolism of the cell.

The importance of STR analysis stems from the observation that while some STRs are constant in their repeat number across entire populations, others are highly variable in repeat number. This variability has been found in all species examined. It has also been shown that not all the different alleles occur at the same frequency in any given species. So for a specified STR some alleles will be more common than others, and the distribution frequently approximates to normal, with very small and very large repeat numbers being rarest. Because this is not always the case we can surmise that some other mechanism than simple random processes are affecting repeat numbers of STRs.

It is the inherent variability of STR repeats which makes them of value [5]. By using a PCR reaction across the STR, given that we know the amount of flanking DNA, the fragment can be accurately sized in terms of the number of repeats present. This means that an STR result is expressed not in terms of base pairs, but as a single figure denoting the number of repeats. STR analysis is one of the simplest PCR processes as analysis of results can be carried out with small agarose-gel electrophoresis systems with a sizing ladder in a second lane. The entire gel can then be stained with ethidium bromide and fluoresced with ultraviolet radiation. This approach requires that, unless widely separated in size, only a single STR is analyzed per experiment. The alternative is to make use of automated systems which can cope with several different STRs at a time by using several different fluorescent markers.

STRs have several important uses for the biologist, among which are personal identification and population studies of animal and plant species. This latter function of STR analysis is especially useful in endangered species which exhibit small residual genetic variation. Since STRs are inherited in a strictly Mendelian fashion, with mutation rates generally less than 1 in

10 000 meiotic events, it is possible to determine the level of inbreeding among a group. STRs can also be used to help breeding programs of endangered species by formulating family trees and aiding captive breeding to optimize outbreeding and reduce suppression of fitness by inbreeding. Mutation of STRs most often results in a change in complete repeat number, but this is not always so; single bases are occasionally added in the middle of an STR. It has been suggested that mutation events only serve to increase the size of the STRs, but simply because reductions have not been seen does not mean they do not occur. A continuous increase in STR size would eventually result in physical restraints on the genome, as well as a reduced ability of the cell to function metabolically.

3.9. Single Nucleotide Polymorphism (SNP) Detection

Single nucleotide polymorphisms (SNPs) are represented in the genome as a single base change in a gene, or other DNA being looked at, although they may be considered less informative than other mutations that have several different variants at a single loci. This is because an SNP will normally be biallelic, that is, it is found in either one of two forms only. However, with SNPs being found at a rate across the human genome varying from one every 500 to 5000 base pairs, depending upon which area of the genome is being looked at, there are potentially up to 3×10^6 SNPs in the human genome. This makes them potentially very useful in a range of areas, especially population structure, pharmacogenetics, and complex genetic traits. These last conditions come under the general heading of polygenic and multifactorial inheritance. Examples of this include a surprising range of conditions that have not generally been thought of in the past to be genetically controlled because the number of genes involved and the interaction of these with the environment made it impossible to analyze them by classical genetic techniques. Such conditions as some forms of heart disease and a person's height can be included in complex genetic traits.

SNP analysis is another PCR-based technique. It can be done in a variety of different ways, but the most straightforward, although not the method with the greatest throughput, involves a modification of the sequencing reaction. By creating an initial PCR reaction which runs for approximately 100 bp across the known SNP, sufficient material can be generated for the determination of the nature of the SNP. The next stage uses a primer immediately adjacent to the SNP, and dideoxynucleotides attached to a fluorescent marker. As soon as the next nucleotide is incorporated, which is in the SNP position, the sequence is terminated. When these products are run out on a gel there will be either one fluorescent band of a single color if the individual is homozygous for that SNP, or two bands of different colors if the individual is heterozygous. Although the two products are essentially the same length in base pairs, the difference in mass with the attached dyes is sufficient for them to be separated. If for any reason they are not separable, then electrophoresis mobility modifiers can be used. The first PCR reaction is necessary to produce enough material for the second-round fluorescence to be detectable.

3.10. Mitochondrial DNA Analysis

Mitochondria are small self-replicating subcellular organelles found in the cytoplasm of all eukaryotes. However, the situation is rather more complicated than that. Mitochondria are the site of oxidative phosphorylation, the energy-producing process in eukaryotes, and although they have their own DNA (mtDNA), only a minority of mitochondrial proteins are actually coded on the mitochondrial genome; most are produced by nuclear DNA. So although self-replicating, they are not autonomous. Generally, mitochondria also lack intron sequences within their genes.

The mitochondrial genome varies from about 15 to 20 kb of double-stranded DNA with approximately 50 gene sequences. Mitochondria are generally thought to be inherited only from the maternal line. This has been shown to be incorrect, but in a way which appears contradictory. In humans most of the mitochondrial genome originates from the maternal side: the ovum. However, it has been demonstrated that the mitochondria of the sperm can, and do, enter the ovum on fertilization. The question then arises as to how much of the mtDNA in the

zygote is of maternal origin and how much is paternal in origin. Generally, most testable mtDNA comes from the maternal side, although the sperm does in fact contribute approximately 100 mtDNA genomes to the ovum on fertilization in mammals. Recent studies have shown that there is a selective destruction of the invading sperm mitochondria soon after fertilization.

Studies of three-dimensional reconstructions of cells by using serial sections and electron microscopy suggest that mitochondria either freely associate with each other or are, in any given cell, single entities. In yeasts the mitochondria fuse and bud off from each other quite frequently. Either way, every cell contain several copies of the mtDNA genome which can be tested [6].

Being generally accepted as maternal in origin, mtDNA has the potential to enable us to track through generations via the maternal pathway of inheritance, rather than paternal. This differentiation has great significance in some applications [7].

Mitochodrial DNA comes in the form of a distorted circle, forming a three dimensional structure as it forms hydrogen bonds with itself. In humans mtDNA is generally about 16 570 base pairs long, but it is not the entire sequence which is interrogated when mtDNA is analyzed. It is a short section of highly variable sequence which is looked at, again by PCR.

The reason mtDNA analysis can be so valuable is that unlike nuclear DNA, mtDNA can be found in robust tissues such as bone shafts and hair shafts, which contain no nuclear DNA at all. These tissues, which are tough and resistant to decay, can harbor useful DNA signatures long after the soft tissues have decayed along with the nuclear DNA and become of no use for determining the origin of a sample.

It is the robust nature of mtDNA and relative ease with which it can be copied in a PCR reaction which makes it ideal for investigating ancient samples such as mummified corpses and bones from ancient burial sites. It has also found favor for determining phylogeographic variability in ancient societies. It is not generally necessary to sequence the entire mtDNA because some areas are intrinsically highly variable. One such is the mitochondrial DNA control region with a high variation rate in the 360 nucleotides at the 5′ end of the sequence and to a lesser degree 200 nucleotides at the 3′ end of the control region. Other methods focus on a section of 780 bases of a variable region, which shows a variation rate of approximately 1 %. By concentrating on 12 sites within this region it becomes easier to produce a result and yet maintain the discriminatory power of the test. Mitochondrial DNA analysis has wide applications from familial identification to questioning species origin in food materials.

3.11. DNA Analysis and Bioinformatics

To analyze DNA from any organism increasingly requires fast and reliable computers. The sheer scope of trying to analyze something of the size of, say, the CF gene which spans over 250 kb can be enormous. It would also be extremely difficult without computers to determine whether a sequence has already been defined amongst all the available sequence data. It is often no longer possible to publish database information as tables or lists, the data are just too much for such a manual system to cope with [8].

When looking for something more precise than just sequence data, perhaps data on mutation rates or sequence variability, it is normal to employ a search engine. Although search engines designed for scientists are slowly becoming available, commercial search engines do not well serve the need of scientists. This is essentially because the software uses crawler programs, which index a page, and then jumps to a linked page and goes through the process again. Commercial web pages often have embedded words in HTML code for their page, or use metatags, words invisible but present on the home page.

The unaudited nature of many web pages, both commercial and scientific is also a potential source of problems. If a laboratory sets up a web page that allows publication of any results they have, two things may occur, neither of which is desirable. The first is that erroneous data may be assumed correct, and the second is that correct and valuable data may be overlooked because of the construction of the web page. This second possibility is all the more distressing since the web has a primary aim of disseminating information as widely as possible. A primary feature of DNA analysis in all its forms is the need for different groups to be able to compare their

results with others working in the same area. Without unified databases this is not easily possible and can almost negate the potential value of having data in an electronic form.

Some large databases containing scientific data have the ability to check new data for consistency and contradiction with what is already available. This is a very valuable facility and underlines the need for well edited and definitive databases. Although there are good starting points for searching a particular area of data or data types, there are also some areas where fractioned data are so widely distributed that it becomes extremely difficult to navigate around the subject. There are for example, more than 150 different sites devoted to human mutations. If it were a simple case of the data being widely distributed, that would make the system complicated enough to deal with, but for many years nomenclature also varied so that it was not always entirely obvious if the same mutation was being referred to. Such was the potential for confusion that the Human Genome Organization (HUGO) has created the Mutation Database Initiative which has produced a consensus on nomenclature and computing conventions so that human mutation data can be cross-referenced with other types of data dealing with both other types of human genetic data and also data from other species. Table 3 lists some useful starting points for data mining.

4. Applications of DNA Analysis

One of the earliest applications of DNA analysis came about indirectly, through the agency of gene mapping [9]. Ordering genes on a chromosome or section of DNA can be carried out with no knowledge of the DNA, gene, or gene product, and such maps have been constructed since the early 1900s by breeding experiments. Creation of such maps became ever more sophisticated to the point where a classical map gave way to data regarding the fundamental chemical nature of DNA.

It is not only by sequencing strands of DNA that we can gain information and knowledge from the genetic code. Indeed, even when the entire human genome has been sequenced, this will only be the very first stage in a long story of research. Knowing a DNA sequence alone gives no indication of the reading frame, the length of the gene, and perhaps more importantly, the nature of the protein product. This last point is particularly significant because even if an amino acid sequence is known from a gene, it may not be clear what the final tertiary structure might be, and consequently what the protein does. A good example of this is the characterization of the cystic fibrosis gene, where the gene was known before the gene product or its action. Also, it may not be necessary to know the entire sequence of a gene to produce a valid analysis of its significance and activity.

Cystic fibrosis (CF) is the commonest severe autosomal recessive condition in Caucasian populations. Approximately 5 % of people are carriers, which results in close to 1 in 2000 live births being affected. For a long time it had been possible to predict the probability of carrier status from a simple family history, but it was not possible to be more precise than that for many years, until it became possible to use linkage analysis to associate specific DNA polymorphisms within a family. This is a process in which a marker, that is, any determinable DNA sequence, is usually transmitted with the disease gene of interest by simply being in very close proximity to it. The logic is therefore that if an individual carries the measurable DNA then they will also carry a defective gene.

The CF gene was thus broadly localized to a specific area on human chromosome 7. Later and progressively detailed analysis of the area produced a candidate gene. This was the first time a gene had been sought and found with no knowledge of the protein product for which it coded. It was only found later, after the protein sequence had been determined, that in 68 % of all CF mutations a single 3 bp deletion of the gene, resulting in the loss of a phenylanaline residue at position 508, was responsible for the problem. This mutation was designated ΔF508: Δ for delete, F for phenyanaline, and 508 for the amino acid position. Since then more than 170 different variant mutations of the gene have been described.

To analyze DNA mutations directly in the CF gene a technique can be used which is widely applicable in many areas of DNA analysis and mutation detection. This is the amplification refractory mutation system (ARMS).

Table 3. Useful starting points for finding genetic data on the World Wide Web

Adress	Type Of Data Contained
http://www.gdb.org/hugo/	Home page of Human Genome Organization
http://www.expasy.ch/sprots/sprot-top.html	Home page of the Swiss-Prot database containing data relating to protein function and variation as well as domain structures
http://www.ncbi.nlm.nih.gov/omin	Home page of the database Online Mendelian Inheritance in Man. This is an annotated listing of traits, both diseases and benign polymorphisms which are inherited in a Mendelian fashion. It also contains data regarding mitochondrial dissorders
http://www.incyte.com./index.shtml	This is a commercial database which contains data relating to sequence data of pathogenic organisms, primarily bacteria and fungi
http://www.ncbi.nlm.nih.gov/genbank/genbankoverview.html	A primary depository for human sequence data, currently containing over 4.5 million sequencies covering approximately 3 billion base pairs
http://www.ncbi.nlm.nih.gov/snp/	A primary depository for SNP data, currently containing in excess of 2.4×10^6 SNPs

4.1. General Principles of ARMS Analysis of DNA

A basic ARMS test requires the use of three primers. These cover the area where a known mutation occurs There is one common primer and two other primers, one which we can refer to as the wild type, that is, the normal, fully functional one, and one which includes the mutation. The length of the PCR products is not in itself important, what is important is the sequence at the end of the mutation detection primers. It is for this reason that the sequence of the gene being studied is known.

If a primer has a mismatch at the 3′ end it is most unlikely to produce a PCR product. So by having one wild-type primer and one mutation primer, two PCR reactions will produce either one product (homozygous normal or homozygous affected,) or two products, indicating heterozygous carrier status. This mismatch specificity is maintained by using a PCR enzyme which does not have a 3′ to 5′ proof-reading ability. If an enzyme is used with this capacity to proof read, no sense can be made of the results. Because the PCR reaction is indicating a specific finding, it is usual to run another, unrelated, PCR reaction in the same tube simply to be certain that the PCR has worked; a form of quality control.

Increased specificity of the reaction can be created by introducing deliberate mismatches close to the 3′ end of the primer. The nearer to the 3′ end the introduced mismatch is, the greater the destabilization, so although specificity may be increased, product yield is likely to decrease.

Earlier, there was no way in which PCR products could be differentiated, and individual reactions had to be carried out. With the advent of fluorescent tagging of products this situation has changed. It is now possible, using only a limited number of different fluorescent wavelengths, to carry out ARMS testing in a single reaction for more than 10 different known mutations. This is done by using a combination of different common primers which results in a series of differently sized products. So the fluorescence might appear as red, blue, yellow, red, green, blue, . . . and so on, essentially the same fluorophore is used several times, but has to come within a sandwich of other fluorophores to be accurately sized.

Another technique, now widely used to construct tests for conditions like CF where there are a large number of possible mutations, all resulting in essentially the same phenotype, is the PCR – oligonucleotide ligation assay (PCR – OLA). In this technique multiplex PCR reactions are followed by a multiplex oligonucleotide ligation reaction which is made up of three probes for each biallelic site to be looked at. One probe is common and is labeled with one of three fluorescent dye markers, and the other two correspond to either the normal or mutant variant at that site. These interrogative probes are attached to various different numbers of electrophoresis mobility markers. These are generally penta(ethylene oxide) units, which can be stacked to give products with the same number of nucleotides but widely differing mobility rates. So when the common probe is ligated to the specific probe it results in a unique com-

bination of fluorescence and mobility. By using only four different fluorescent dyes, three for the probes and one as a lane size marker, it is possible to size a large number of different products in the same lane and determine the presence or absence of mutant or normal alleles at each site within the gene.

4.2. Analysis of Dynamic Mutations

A group of human conditions has caused considerable consternation to those trying to define them clearly. These disorders are characterized by dynamic mutations that result in familial transmission, but under some conditions not necessarily in a clearly Mendelian fashion. In those that are clearly inherited in a dominant–recessive way, the age of onset may be extremely variable. This latter state could sometimes be mistakenly associated with closely linked polygenic inheritance or environmental factors. Although relatively few of these conditions have so far been described, they are very important, either because of their frequency or severity. These include Fragile X syndrome, Huntington's Chorea, and Myotonic dystrophy, in which normal dominant inheritance of the condition is modified by a dynamic mutation in the 3′-untranslated region of the gene for myotonin protein kinase.

One of the commonest genetic diseases and certainly the commonest form of familial mental retardation is Fragile X syndrome, affecting approximately 1 in 2500 children. The condition gets its name from the fragile site found at the end of the X chromosome. For many years the only way of diagnosing this condition was cytogenetically, which was inaccurate, time-consuming, and very expensive. Cytogenetic techniques gradually gave way to rather cumbersome pre-screening using molecular methods. Now, however, it is possible to gain a much clearer picture by using PCR in a technique which is equally applicable to other dynamic mutations.

Dynamic mutations are those which are represented not by a change in base composition, but a progressive lengthening of a sequence. This takes the form of progressive expansion of a trinucleotide repeat. In the case of Fragile X it is an increase in the number of polymorphic CCG repeats in the 5′-untranslated region of the gene. The number of repeats varies from about 6 to 55 in normal X chromosomes, but towards the upper limit of 55 the copy number becomes unstable between generations. With over 55 repeats the chromosome is said to have a premutation, and the instability increases dramatically, that is, a dynamic mutation is one where once changed the DNA has an altered risk of further change. When the number of copies exceeds 230 it is a full mutation. It is therefore useful to know how many copies are present in the premutation state. This is done by PCR reactions across the area of the trinucleotide repeat. With very large numbers of repeats in these conditions it may sometimes be necessary to produce products which are more than 2.5 kb long. This can be extremely difficult as the probability of damage to target DNA is proportional to its length. Hence, it is important to extract the DNA mildly and avoid shearing it when using pipettes. An internal control can be used to check the voracity of the PCR reaction. Figure 7 shows the results of sizing the PCR products from three normal individuals, homozygous, heterozygous, and what is commonly referred to as the N+1 heterozygote where there is only one repeat difference between the two alleles. The other peaks represent the ApoE check on PCR and the sizing ladder used to measure the number of repeats.

4.3. Using DNA Analysis to Determine Sex

Humans differ chromosomally between the sexes in that females are the homogametic sex having two identical X chromosomes and males are the heterogametic sex having one X and one Y chromosome. There is a gene which codes for a protein associated with development of tooth enamel called amelogenin. By running a PCR across part of the gene two fragments are generated, one of 212 bases from the X chromosome and the other of 218 bases from the inactive gene on the Y chromosome. These are created by the same primers, so by running the products on a gel there will be two bands from a male and only one band from a female. Although the gene on the Y chromosome is inert, it does not seem to be susceptible to alteration, so this test is now

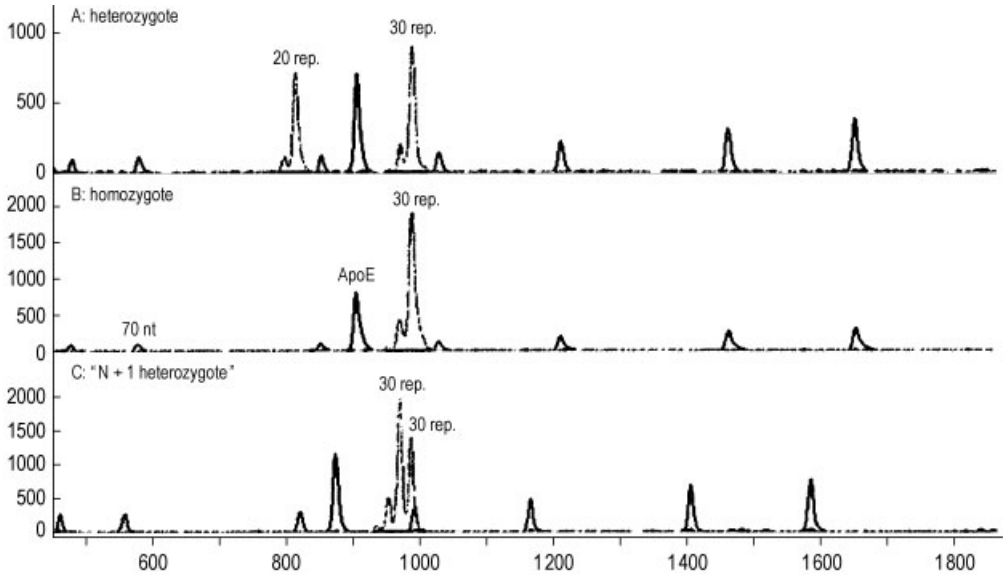

Figure 7. The automated output from measurement of repeat numbers in Fragile X analysis. The three traces represent heterozygote, homozygote, and N+1 heterozygote. The peak marked ApoE is a check that the PCR has run adequately, and the other peaks are the sizing ladder. This technique is readily extended to other conditions involving this sort of dynamic repeat expansion.
Image courtesy of PE Applied Biosystems

routinely used on forensic samples to determine the sex of the originator of the sample.

Birds are quite different. Here it is the males which are homogametic and females heterogametic. Because of this difference the sex chromosomes are designated Z and W, with males being ZZ and females ZW. There is a highly conserved sequence on the sex chromosomes of birds which differ in only a minor way. By running a PCR reaction across the highly conserved region, enough product is produced for the second stage. The difference between the two is a restriction site which can be cut by the restriction endonuclease *DdeI*. This operates by cutting at the sequence 5′CTNAG3′. The W chromosome produces a 73 bp fragment and the Z chromosome a 104 bp fragment, so there will be either one or two bands on a gel depending on the sex of the bird being tested. Because of the highly conserved nature of the tested sequence, it has been possible to use it in virtually any bird group other than the Ratites (emus, ostriches, etc). Since both sexes of many bird species appear identical, it is important to be able to determine the sex of the birds in captivity, so that breeding programs can make the best use of stocks [10].

4.4. Methods of Personal Identification

Many different methods of personal identification have been developed in the last 25 years on the basis of DNA analysis. The more recently developed ones require more automated equipment but less human intervention, making interpretation of the results far less subjective. The first systems required relatively large quantities of material, usually blood, but as technology progressed so the amount of starting material has decreased, and it is now theoretically possible to determine with a good degree of certainty the generator of a sample. This is especially important in two areas of law: criminal cases and paternity disputes. Although specific methods are used in such cases, this is more for continuity of service and agreement of acceptable methods, rather than one being better than another.

All methods which can discriminate between two sequences can in principle be used to discriminate between two individuals, but the level at which they can be relied upon varies enormously. This reliability can stem just from factors such as inadequate discriminatory power of the test or subjective interpretation of the results, this latter situation being more important in ear-

lier systems that were implemented rather than later ones.

The first methods of personal identification involved the use of mulitlocus (MLP) and single-locus probes (SLP). A sample of native DNA was digested with an enzyme and then run out on a gel. This could then be probed to give several bands per individual in the case of MLPs, or one or two bands in the case of an SLP analysis. The major problem with these techniques is the size of the fragments which are probed. As the measured DNA can run to several kilobases, separation inevitably involves interpretation and agreement on methods of being able to say that a large band does in fact only comprise DNA of a particular type. It is possible that a band that appears to represent 14 kb may in fact be 14.1 kb and therefore different, but not separable, on this scale. It is these sorts of questions which helped to move techniques of personal identification along.

Although the ultimate personal identification method would be to sequence an individuals genome, this is at present not a practical possibility. However, two good methods are available and are routinely used, namely, STR analysis and SNP. Both suffer from an intrinsically low power of discrimination. For example, while any given STR may have 10 or more possible alleles, they will not all be equally common, so a single STR may have a frequency within the population of 1 in 10, that is, 1 in 10 of the population would be expected to have that allele. Some may be even more common. To get around this, a panel of STRs is used, either singly or in one PCR reaction mixture. By multiplying the individual frequencies it is possible to generate a composite value for the probability of finding the total combination of STRs by chance, and therefore a likelihood ratio can be generated for that profile coming from a specific individual. With SNP analysis this procedure is taken slightly further, because for any given SNP the number of possible variants is both smaller and fixed, but the method is essentially the same: production of an accumulated probability by analysis of several different SNPs.

When comparing, say, material from a body with a reference sample, mutation rates are unimportant. But in questions of paternity mutation from one generation to another is a natural consequence of genetic variation. This is one reason why the number of sites looked at is usually much bigger than in situations where a sample is compared directly with an individual [11]. Another reason for this increase is that it gives a greater degree of certainty in the production of a family tree, because only half the loci in the genome of a child comes from the father.

A specialized, and very limited, method which can be employed in personal identification is mtDNA analysis. This procedure is limited because it can link a child to a mother but not to a father, and siblings can not be distinguished by means of mtDNA. It is useful, however, where remains have been found and only mtDNA has survived, for example, after burial or fire. Under these circumstances mtDNA analysis can be used to reunite the remains with the right family.

4.5. Bacterial Contamination of Water Supplies

A practical application of DNA analysis is in detecting coliform bacteria in water supplies. The same techniques can be extended to check for pollution of surface water and beaches by sewage. The importance of this is due to the range of pathogenic microorganisms that occur in fecal material: *Salmonella* (causing typhoid fever), *Shigella* (giving rise to dysentery), and *Vibrio cholerae* (the causal agent of cholera). Another important contaminant is *Legionella pneumophila*, often present in air conditioning units and the cause of legionnaires disease.

Although it has long been possible to detect these organisms, the traditional methods of microscopy and viable plate count are not necessary quantitative methods. Sensitivity is also inadequate with these systems, being approximately 10^4 cells per milliliter, which is far too low for sensitive issues of environmental contamination. A better and more sensitive way of assessing both the presence and quantity of these organisms is to use a PCR reaction coupled with a suitable probe. Sensitivity is now orders of magnitude greater than previously achievable, and detection is possible far beyond the limits required by public health regulations. Thus it is possible to detect as few as 1 – 10 cells per milliliter, which equates to between 1 and 10 fg of DNA depending on the target organism.

For these environmentally important species a panel of very precise probes are used after DNA extraction and PCR of target DNA. The very high specificity of this technique is essential, as there are frequently several different species in a sample, most of which will be benign. Because the detection is based on a presence/absence assay, the PCR products are not run out on a gel, but dot blotted onto a suitable membrane, usually nylon or one of the many proprietary brands available. Once immobilized onto the membrane a probe is used which is labeled at the 5′ end. These dot blots are then probed with the complementary sequence to the target. For *Legionella pneumophila* the probe is based on the macrophage infecting protein (*mip*), and for total coliform bacteria it is based upon a section of the *lacZ* gene of *Escherichia coli*.

4.6. Adulteration of Food Stuffs

Food labeling is not always as accurate as might be supposed, and sometimes accuracy must be checked. Questions which can be asked are, for example, whether a tin of tuna really contains tuna, and are meat products derived from the species that is stated on the label. It is even possible to determine whether pasta is correctly made from hard Durum wheat or from bread wheat.

Given the importance of checking foods for adulteration this is one of the most rapidly developing areas of applied DNA analysis. The constraints are similar to environmental testing, such that it is essential to have an unequivocal standard against which to measure the result. This may seem self-evident, but if you are trying to decide whether a food is what it says it is, there is little point in comparing your results against a standard for which you do not have the provenance, or chain of custody. Mistakes along these lines have happened.

Generally, the methods of checking for authenticity of species origin are based on DNA targets within the cytochrome B gene. This is located on the mitochondrial DNA and has a high species specificity. Although autoclaving is usually regarded as an adequate decontaminating procedure for material being used in DNA analysis involving PCR, this may not necessarily be so. For example, it is possible to determine the species of animal from which a pet food is produced, even though it has been rigorously cooked at high temperature and pressure. Similarly the checking of canned tuna for authenticity is important. This is because "tuna" covers a range of fish in several different genera, so to be sure that a can labelled "skipjack", really is of the genus *Katsuwonus* rather than *Thunnus* (tunny) or *Germo* (the albacore) it becomes necessary to use DNA analysis. In these cases, as with any cooked product, it is generally possible to produce a repeatable result, unless the food stuff has been cooked at high temperature in an acidic liquor, after which results are unreliable and frequently nonexistent.

Whenever food adulteration is suspected, it should be remembered that PCR is a technique of extraordinary sensitivity. Any result must be quantified, and the possibility of accidental contamination must always be considered.

5. References

1. B. Budowle, R. Chakraborty, A. M. Giusti, A. J. Eisenberg, R. C. Allen, *Am. J. Hum. Genetics* **48** (1991) 137 – 144.
2. C. M. Hearne, S. Ghsoh, J. A. Todd, *TIG* **8** (1992) 288 – 294.
3. L. Kricka (ed.): *Nonisotopic DNA Probe Techniques,* Academic Press, London 1992.
4. K. M. Sullivan, S. Pope, P. Gill, J. M. Robertson, *PCR Methods and Applications* **2** (1992) 34 – 40.
5. C. P. Kimpton, N. J. Oldroyd, S. K. Watson, R. Frazier, P. E. Johnson, et al., *Electrophoresis* **17** (1996) 1283 – 1293.
6. S. Anderson, A. T. Bankier, B. G. Barrell, M. H. L. deBruijn, A. R. Coulson, et al., *Nature (London)* **290** (1981) 457 – 465.
7. E. Hagelberg, J. B. Clegg, *Proc. R. Soc. London* **B 252** (1993) 163 – 170.
8. S. M. Maurer, R. B. Firestone, C. R. Scriver, *Nature (London)* **405** (2000) 117 – 120.
9. D. Chadwick, G. Cardew (eds.): *Variation in the Human Genome,* Ciba Foundation Symposium 197, John Wiley and Sons, Chichester 1996.
10. R. Griffiths, B. Tiwari, *Nature (London)* **375** 454.
11. R. L. Alford, H. A. Hammond, I. Coto, C. T. Caskey, *Am. J. Hum. Genet.* **55** (1994) 190 – 195.

Temperature Measurement

MANFRED SCHLEICHER, JUMO Mess- und Regeltechnik, Fulda, Federal Republic of Germany

1. Definition of Temperature 1
2. History of Thermometers 1
3. Temperature Scales 1
4. Temperature Measurement 3
5. References 11

1. Definition of Temperature

Heat is a measure of the energy in a body due to the irregular movement of its molecules or atoms. Temperature is a measure of the internal energy of a body. The internal energy contained in a body is the product of its temperature, its mass, and its specific heat capacity.

2. History of Thermometers

The first physical thermometer, constructed in 1592 by GALILEO utilized the thermal expansion of air. A glass column closed at one end was placed with its open end into a vessel containing a liquid. Variations in temperature altered the air volume, causing the liquid level in the glass tube to rise and fall. The liquid level was therefore a measure of temperature. SANCTORIUS later added a scale to the thermometer. The disadvantage of this type of thermometer was that its reading varied with atmospheric pressure.

In 1660 the first liquid-in-glass thermometer that was not affected by external atmospheric pressure was manufactured for the Duke of Tuscany, Ferdinand II. A glass vessel was filled with a liquid and fitted with a capillary closed at one end. The scale consisted of a maximum value (the highest summer temperature) and a minimum value (the coldest snow). The thermometer was calibrated by RENALDINI in 1694 at the ice point and at the boiling point. The liquid-in-glass thermometers still in use today were developed in 1724 by the Danzig glassblower FAHRENHEIT.

The first Bourdon spring thermometers date from the year 1879. The first liquid-filled Bourdon spring thermometers with a capillary in the form in use today were reported by DREYER-BÜRKNER in 1882.

Bimetal thermometers were invented around 1800 by soldering together two metals with different expansion coefficients. Metal expansion thermometers were manufactured around 1860.

SEEBECK discovered thermoelectricity in 1821; its use for measurement was developed by LE CHATELIER in 1887 and by C. BARUS in 1889. The platinum thermometer – a resistance varying with temperature – was utilized by SIEMENS in 1871. It was further developed by CALLENDAR in 1886 into a precision thermometer, also for measuring higher temperatures.

A principle for noncontacting temperature measurement was proposed in 1880 by CROVA; the radiation densities at two different wavelength were utilized to measure temperature. On this basis LE CHATELIER produced in 1892 the first usable spectral pyrometer.

3. Temperature Scales

As early as 1694 RENALDINI used the melting point and boiling point of water for the calibration of temperatures, since these points are readily reproducible. FAHRENHEIT used as fixed point an ammonium chloride freezing solution, to which he assigned the value 0°, and placed normal human body temperature at 96°.

Around the same time Frenchman Réaumur set the ice point at 0° and the boiling point at 80°. In 1742 the Swedish astronomer CELSIUS introduced the division into 100 degrees, with 0° as the melting point and 100° as the boiling point of water.

In 1887, the hydrogen scale of the Comité International des Poids et Mesures was agreed internationally, and it was decided that a hydrogen thermometer would be the fundamental measuring instrument.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.e26_e02

As, according to GAY-LUSSAC (1802), in an ideal gas, the product of volume and pressure increases by 1/273.15 of its initial value per degree, the temperature scale of a gas thermometer can be extended below 0 °C and above 100 °C. According to GAY-LUSSAC, for ideal gases:

$$\frac{(pV)_1}{(pV)_0} = 1 + \alpha t = \alpha\left(t + \frac{1}{\alpha}\right) = \alpha T \tag{3.1}$$

where $(pV)_0$ is the value of the product pV at $t = 0$ °C, $\alpha = 1/273.15$ K and $T = t + 273.15$ K, and T is the temperature relative to absolute zero ($t_a = -273.15$ °C). It is defined as the absolute temperature and was introduced in 1851 by KELVIN. Absolute temperatures are measured in kelvin (K).

These gas thermometers, filled with hydrogen or helium, have their limitations, as even these gases no longer behave as ideal gases at lower temperatures.

The first reproducible temperature scale whose validity was independent of the temperature and of properties of substances was established in 1852 by KELVIN. It is based only on the Second Law of Thermodynamics and is referred to as the thermodynamic temperature scale. Kelvin established for the Carnot cycle a relationship between the initial temperature, the energy flow, and the resulting final temperature. From this relationship, the temperature T and hence also the temperature scale is unequivocally defined as the SI unit of thermodynamic temperature (kelvin). 1 K is 1/273.16 of the thermodynamic temperature of the triple point of water which can be realized with very low measurement uncertainty. The kelvin is also the unit for temperature differences and temperature intervals.

Celsius temperature is defined as the difference between a given Kelvin temperature T and the temperature $T_0 = 0$ K. Temperatures on the thermodynamic temperature scale that correspond to degrees Fahrenheit are expressed in degrees Rankine (°R).

Conversions between the temperature scales can be calculated with the aid of the following equations:

$$T_\mathrm{K} = 273.15 + t_\mathrm{C} = 255.37 + \frac{5}{9}t_\mathrm{F} = \frac{5}{9}T_\mathrm{R} \tag{3.2}$$

$$t_\mathrm{C} = T_\mathrm{K} - 273.15 = \frac{5}{9}(t_\mathrm{F} - 32) = \frac{5}{9}T_\mathrm{R} - 273.15 \tag{3.3}$$

$$T_\mathrm{R} = \frac{9}{5}t_\mathrm{C} + 491.67 = \frac{9}{5}t_\mathrm{K} = t_\mathrm{F} + 459.67 \tag{3.4}$$

$$t_\mathrm{F} = \frac{9}{5}t_\mathrm{C} + 32 = \frac{9}{5}T_\mathrm{K} - 459.67 = T_\mathrm{R} - 459.67 \tag{3.5}$$

where T_K = temperature in K, t_C = temperature in °C, T_R = temperature in °R, and t_f = temperature in °F.

The realization of the thermodynamic temperature scale by means of the Carnot cycle is experimentally laborious. However, the thermodynamic temperature can be determined from all measurement processes that can be derived from the Second Law of Thermodynamics, for example, the general gas equation:

$$pV = nRT_\mathrm{K} \tag{3.6}$$

where p = gas pressure, V = gas volume, n = amount of substance, and $R = 8.314\ \mathrm{J\,mol^{-1}\,K^{-1}}$ (molar or general gas constant).

This expresses a direct relationship between the pressure, the volume and the temperature of an ideal gas. A temperature can therefore be measured by the pressure of a gas. In such a process, no substance-dependent quantities or conversion factors are necessary.

In metrological institutes, the thermodynamic temperature is usually measured with the aid of gas pressure. However, as this measurement method is also relatively difficult and expensive, it was agreed in 1927 to create a practical temperature scale which mirrors the thermodynamic temperature scale as closely as possible. Fixed points (solidification points, triple points, melting points, etc.) of substances are used, and the corresponding thermodynamic temperatures are measured. Based on a large number of comparative measurements performed in the national metrological institutes, the values are then fixed.

By the end of 1989, the International Practical Temperature Scale of 1968, the IPTS-68, had become generally accepted. In 1990, a new scale came into force, the International Temperature Scale, ITS-90. This new scale was necessary because a large number of measurements in various

laboratories in the world showed inaccuracies in the determination of the fixed-point temperatures. The defined fixed points of the ITS-90 are given in Table 1.

Table 1. Fixed points of the ITS-90

	T_{90}, K	t_{90}, °C
Vapor pressure He*	3 to 5	− 270.15 to − 268.15
Triple point H	13.8033	− 259.3467
Vapor pressure H*	17	− 256.15
Vapor pressure H*	20.3	− 252.85
Triple point Ne	24.5561	− 248.5939
Triple point O_2	54.3584	− 218.7916
Triple point Ar	83.8058	− 189.3442
Triple point Hg	234.3156	− 38.8344
Triple point H_2O	273.16	0.01
Melting point Ga	302.9146	29.7646
Solidification point In	429.7485	156.5985
Solidification point Sn	505.078	231.928
Solidification point Zn	692.677	419.527
Solidification point Al	933.473	660.78
Solidification point Ag	1234.93	961.78
Solidification point Au	1337.33	1064.18
Solidification point Cu	1357.77	1084.62

* Calculation of temperature from the vapor pressure/temperature relationship.

For the intermediate values, the temperature scale is defined with the aid of interpolating measuring equipment, for example in the range from − 259 to 961 °C by the platinum resistance thermometer. The thermometer must be constructed in accordance with the requirements of ITS-90 and must be calibrated with the aid of the specified fixed points.

4. Temperature Measurement

Liquid-in-Glass Thermometers. In the liquid-in-glass thermometer the thermal expansion of a liquid is used for determining the temperature. The liquid is placed in a thin-walled glass vessel which is attached to a capillary. The capillary carries a scale from which the temperature can be read. Glass thermometers are used today in domestic and laboratory applications.

Mechanical Dial Thermometers. Mechanical thermometers convert the thermal expansion of gases, liquids, or metals into a change in the position of a pointer. The temperature scale is positioned locally but may also be located a few metres from the measurement point. Compared with glass thermometers, mechanical thermometers are unaffected by vibrations and contain no substances which could endanger the environment. The thermometers operate without any external energy, permit the switching of limit contacts, and subject to the appropriate approval can be used in explosion-hazard areas. They offer long-term stability and, unlike electrical thermometers, are unaffected by electromagnetic influences.

The most widely used mechanical thermometers are the gas-pressure and the liquid-filled Bourdon thermometers and bimetal thermometers.

Gas-Pressure Bourdon Thermometer. The general equation of state of an ideal gas is:

$$\frac{pV}{T_{\mathrm{K}}} = \text{const.} \tag{4.1}$$

where p is the pressure of the gas, V the volume, and T_{K} the temperature of the gas in kelvin.

If the temperature of a gas contained in a closed system is increased (e.g., in a cylinder with a sliding piston), the product pV increases correspondingly. Assuming an isobaric process (temperature change at constant pressure) the following relationship applies to the volume:

$$V_{\mathrm{T_{new}}} = \frac{T_{\mathrm{new}}}{T_{\mathrm{old}}} V_{T_{\mathrm{old}}} \tag{4.2}$$

where $V_{\mathrm{T_{new}}}$ is the volume at the new temperature, T_{new} and $V_{\mathrm{T_{old}}}$ = volume at the old temperature T_{old} in K.

Replacing T_{new} by $T_{\mathrm{old}} + \Delta T$ (ΔT = temperature increase of the gas) gives

$$V_{\mathrm{T_{new}}} = \left(1 + \frac{\Delta T}{T_{\mathrm{old}}}\right) V_{T_{\mathrm{old}}} \tag{4.3}$$

Equation (4.3) describes in idealized form the process in a gas-pressure thermometer: the volume of a gas (usually nitrogen) increases in proportion to the temperature.

The gas-pressure Bourdon thermometer (Fig. 1) consists of the actual temperature probe, which is a metal vessel containing the gas volume. As the gas expands with increasing temperature it passes through the capillary to the elastic measuring element (Bourdon tube). The

inflowing gas causes the Bourdon tube to expand. This movement is transferred to a pointer, the position of which indicates the temperature on a scale.

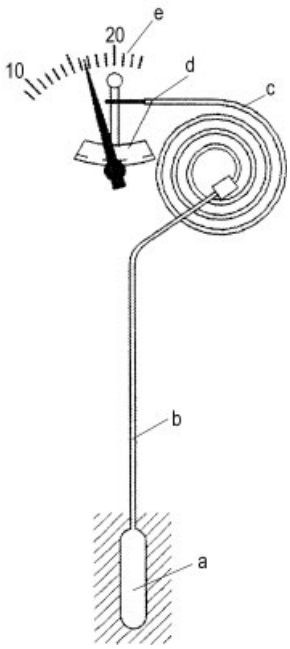

Figure 1. Gas-pressure Bourdon thermometer
a) Temperature probe; b) Capillary; c) Bourdon tube; d) Pointer; e) Scale

The measuring range of gas-pressure Bourdon thermometers is between −100 and +600 °C, their accuracy is Class 1 (according to EN 13190). In practice the upper limit of the range should not be too low, otherwise excessive pressurization of the gas is required.

Part of the gas is always inside the capillary rather than the measuring element and is therefore not exposed to the temperature of the measurement location. The ambient temperature influences this part of the gas and therefore the result of the measurement. The ambient-temperature effect, expressed as percentage of range for gas-filled thermometers, referred to 23 °C is 0.1 %/K (max.) for housing and 0.005 % m^{-1} K^{-1} (max.) for the capillary.

Liquid-filled Bourdon Thermometer. The volume change of a liquid with increasing temperature is described by the following equation:

$$V_{T_{new}} = (1 + \gamma \Delta T)\, V_{T_{old}} \tag{4.4}$$

where $V_{T_{new}}$ is the volume after temperature change by ΔT in K, $V_{T_{old}}$ the volume before temperature change, and γ the volumetric expansion coefficient in K^{-1}.

The relationship indicated in Equation (4.4) is an idealized description of the process in a liquid-filled Bourdon thermometer: a liquid volume (petroleum or Diphyl KT) increases proportionately with temperature. The construction of the thermometer is similar to that of the gas-pressure Bourdon thermometer. Liquid-filled Bourdon thermometers cover the range −40 to +350 °C; their accuracy is Class 1.5 or 2 (according to EN 13190).

The ambient-temperature error, expressed in percentage of range, for liquid-filled Bourdon thermometers, referred to 23 °C is 0.15 %/K (max.) for the housing, and 0.015 % m^{-1} K^{-1} for the capillary.

Bimetal Thermometers. Metals expand with increasing temperature, like most solid bodies. The expansion depends on the temperature and on the type of metal:

$$l_{new} = l_{old}\,(1 + \alpha\, \Delta T_K) \tag{4.5}$$

where l_{new} is the length of the metal after temperature increase by ΔT_K, l_{old} the initial length of the metal, and α the linear expansion coefficient of the metal.

If two metals with different linear expansion coefficients are combined into a bimetallic strip and exposed to a temperature rise, the strip deforms into a curve. This effect is utilized in the bimetal thermometer.

The temperature-sensitive component consists of a spiral or coiled spring which is wound from a bimetallic strip. A temperature change causes the spring to close up or to unwind. In the simplest case the spring is fixed on the outside and carries a light plastic pointer on its inner end. Other arrangements with an immersion tube permit some separation between sensing element and indicator; the immersion tube contains a spindle which is rotated by the bimetallic coil and transmits this rotation to a pointer at the end of the immersion tube.

The range of these thermometers is between −80 and +500 °C. Their accuracy conforms to Class 1 or 1.5 (according to EN 13190).

Electrical Thermometers. In industrial processes the measurement site is often remote from the point of indication. Often, the temperature reading must be processed in controllers or recorders. Such applications are unsuitable for directly indicating thermometers and require types which convert the temperature into an electrical signal. The most widely used sensors of this type are the platinum resistance thermometer and the thermocouple.

Platinum Resistance Thermometer. The electrical conductivity of metals is based on the mobility of conduction electrons. The movement of electrons is, among other parameters, dependent on temperature. Increasing the temperature causes the oscillations of the atoms to increase, raising the overall resistance of the metal. This effect is utilized in the platinum resistance thermometer (Fig. 2).

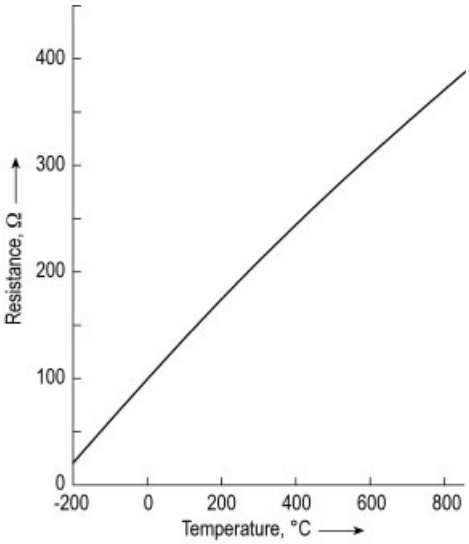

Figure 2. Characteristic of a platinum sensor Pt100 (EN 60751)

Platinum resistance thermometers are defined by EN 60751 according to different tolerance classes (Fig. 3).

Temperatures can also be measured with resistors whose variation with temperature is not standardized. They involve a very inconvenient calibration of the electronics and are not fully interchangeable, so that they do not play an important part in industrial temperature measurement.

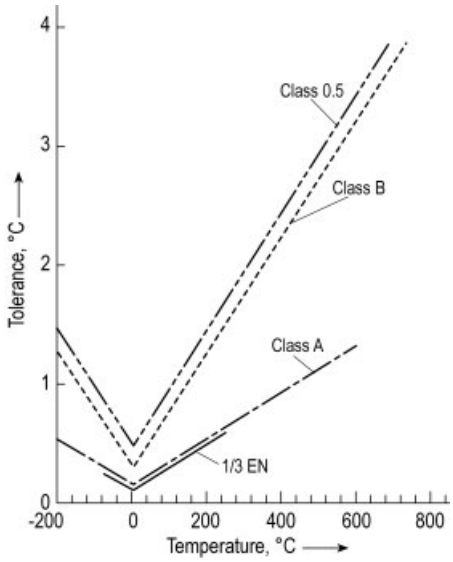

Figure 3. Tolerances of the platinum resistance thermometers (EN 60751)

In the platinum resistance thermometer, a defined and constant current I is passed through the element, and the resulting voltage is used to deduce the resistance and therefore the temperature. This arrangement forms a connection in a *two-wire circuit*. The measured voltage is given by:

$$U = (R_{\mathrm{Pt}} + 2\,R_{\mathrm{L}})\,I \tag{4.6}$$

where U is the measured voltage, R_{PT} the resistance, R_{L} the lead resistance, and I the current.

The measured voltage, the measure of temperature, includes twice the lead resistance. Even where this is allowed for in the arrangement of the measuring system it is not possible with this type of circuit to compensate errors due to varying lead resistance (temperature changes of the leads).

For this reason the *three-wire circuit* (Fig. 4) has an additional lead connected to one end of the measuring resistor. The measurement consists of the voltage at this point (potential against 0 V), and the total voltage across the sum of measurement resistor and the lead resistances.

The total voltage in the evaluation electronics is given by:

$$U = (R_{\mathrm{Pt}} + 2\,R_{\mathrm{L}})\,I \tag{4.7}$$

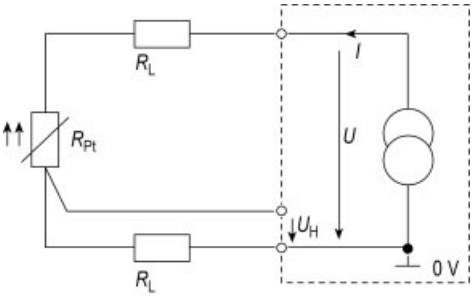

Figure 4. Platinum resistor in a three-wire circuit arrangement

The voltage at the third lead against 0 V amounts to:

$$U_H = I R_L \tag{4.8}$$

U_H = voltage against 0 V at the third lead.

To obtain the actual voltage at the sensor the electronics deducts $2\,U_H$ from U.

The *four-wire circuit* involves two additional leads which are used to measure the voltage directly at the measuring element. Variations in the lead resistances then have no influence on the measurement.

When using resistance thermometers in a two-wire circuit, any change in the temperature of the leads results in alterations in the measured temperature since the altered lead resistance forms part of the measured resistance. If this influence is too large it is advisable to operate the resistance thermometer in a three- or four-wire circuit.

The resistance of the platinum sensor has certain nonlinearities with temperature. Controllers and recorders manufactured by JUMO therefore incorporate linearization tables which can be activated when a platinum sensor is connected.

Measuring resistors are supplied in different designs:

The *ceramic temperature sensor* has a platinum coil of 0.010 to 0.035 mm diameter wire arranged in two bores of a ceramic tube. The bores are filled with alumina powder to secure the winding and to improve heat transfer. The winding has no direct contact with the housing. On changes in temperature the platinum winding can expand freely and experiences only very small forces, and this results in a sensor with a high long-term stability. The sensor has a very wide temperature range of − 200 to + 800 °C (sometimes up to 1000 °C) but is very expensive due to the large amount of manual work involved. Dimensions: diameter 0.9 – 5.0 mm, length 10 – 35 mm.

In *glass temperature sensors* the platinum wire is wound on a glass former and hermetically sealed in glass, permitting direct use in a fluid. The sensor is very resistant to shock and can be used at temperatures between − 200 and + 400 °C; it is also very expensive because of the large amount of manual work. Dimensions: diameter 0.9 – 5.0 mm, length 7 – 55 mm.

The *foil temperature sensor* consists of two polyimide foils enclosing a platinum wire. This flexible temperature sensor is used for measurement on surfaces, pipes, and tools in the temperature range − 80 to + 180 °C. The sensor has a high dielectric strength (up to 2 kV), but because of the mechanical loading its stability is less than that of wire-wound temperature sensors. The large amount of manual work again results in a high price. Dimensions: length 50 mm, width 21 mm, thickness 0.4 mm.

The sensor with the largest market share is certainly the *thin-film resistor* (Fig. 5). The sensor has a permitted temperature range of − 70 to + 600 °C. Produced by large-scale manufacture, the low-cost thin-film measurement resistor has a wide range of applications (laboratory, industrial measurement and control, medicine, white goods, cars, etc.). Dimensions: length 2.3 – 10 mm, width 1.5 – 2 mm.

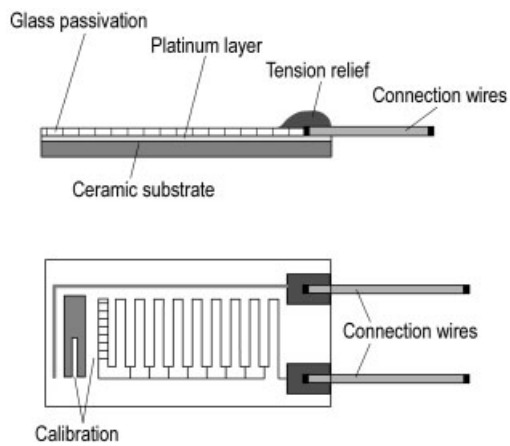

Figure 5. Thin-film measurement resistor

Long-term stability of resistance thermometers is largely influenced by the conditions of use and structural features. Temperature cycling between upper and lower temperature limits involves much more stress on all components than a steady load. Ceramic sensors exhibit very stable long-term behavior, since mechanical coupling is limited by embedding the wire winding in powder.

With glass and thin-film sensors there is very close coupling between the temperature-dependent platinum and the glass or ceramic substrate. This produces changes in the crystal structure of the platinum and results in changes in its resistance. In the manufacture of glass sensors this problem is met by selecting a glass with approximately the same temperature expansion coefficient as platinum. In addition the sensors are artificially aged. Intensive research and development has led to the development of thin-film resistors which also offer good long-term stability.

A further error in temperature indication may be caused by insufficient resistance of the insulation. Particularly in ceramic sensors, the insulation resistance decreases with increasing temperature above 600 °C. Insufficient insulation resistance may also be caused by ingress of moisture, which must be prevented by appropriate measures.

Temperature measurement with a resistance thermometer involves the passage of a current; this produces heating of the sensor and therefore an increase in the indicated temperature. This self-heat depends on the measurement current, the basic resistance of the temperature sensor, and on how readily heat can be dissipated.

The connecting wires of resistance thermometers often consist of silver as internal connections which are extended by copper or nickel. The thermal voltages produced at these changeover points only cancel out if they are at the same temperature. With unbalanced voltages the resulting thermal voltage is interpreted by the evaluation electronics as a voltage drop, resulting in a measuring error. These thermal voltages can be completely suppressed by operating the resistance thermometer with alternating current instead of direct current.

The Thermocouple. If a metal is heated at one end, the heat supplied increases the mean velocity of the free electrons. The electrons drift from the warmer to the cooler region and form a negative charge there. A voltage is generated between the cold and the warm regions inside the metal body. To measure this voltage a second conductor of the same material must be placed in contact with the warm end. A similar charge is produced inside the second conductor so that the potential difference is 0 V. If, however, a different metal is used whose electrical conductivity is poorer than that of the first metal, the reduced number of electrons at the cold end results in a lower negative potential. The difference between the two potentials can be measured as a voltage between the two ends. The negative pole of the thermocouple appears at the metal that has the higher electrical conductivity.

From the large number of metal combinations a few have been selected in EN 60584 and standardized with respect to their voltage calibration and permitted tolerances. Such thermocouples are therefore fully interchangeable.

EN 60584, Part 2, defines the tolerances in three classes. Classes 1 and 2 apply to the range above − 40 °C while Class 3 is essentially for negative temperatures down to − 200 °C.

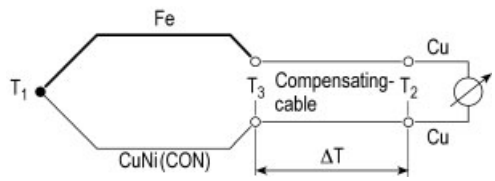

Figure 6. The thermocouple measurement chain

Figure 6 shows the connections to a thermocouple. Thermocouples are frequently located at some distance from the measuring device; since the thermocouple is made from expensive materials, the connection to the measuring instrument is made through a compensating cable which has the same thermoelectric properties as the actual thermocouple. The voltages at the changeover points compensating cable Fe – Cu and Cu – compensating cable CuNi are added together to form the thermal voltage which would be produced by a direct connection of the two compensating leads. The voltameter therefore measures the thermal voltage of the Fe – CuNi thermocouple at the measuring point (T_1) minus the thermal voltage at the changeover to copper (T_2). The voltage measured by the instrument is

a measure of the temperature difference between T_1 and T_2; the electronics must therefore measure the temperature T_2 and add it to the temperature difference determined from the thermal voltage.

Foreign atoms which become absorbed in the thermocouple material influence the binding force of the electrons and produce a change in its thermoelectric properties; the thermocouple calibration starts to drift. To limit this effect, as well as other external influences the thermocouple is placed inside a protection tube or is embedded together with an oxidic insulating material inside a metal sheath (mineral-insulated thermocouple).

Selection criteria for different thermocouples, principally the use temperature and the magnitude of the thermal voltage, are given in the following:

Cu – CuNi	not widely used
(Cu – Con)	oxidizes already above 200 °C
	copper: high thermal conductivity favours heat-conduction errors
Fe – CuNi	widely used (low cost)
(Fe – Con)	used in the low and medium temperature range
	iron becomes brittle a sulfur-containing atmospheres
	alson used as mineral-insulated thermocouple
NiCr – CuNi	used in the low-temperature range because of its large thermal voltage
(NiCr-Con)	widely used in the USA
Ni – CrNi	used above 600 °C
	good long-term stability (ten times better than Fe – CuNi)
	maximum temperature as mineral -insulated thermocouple: 1150 °C
NiCrSi – NiSi	can be used up to 1300 °C
	as mineral-insulated thermocouple with suitable sheath material also stable at high temperatures
	has only been on the market for a few years
Noble metals	used only at temperatures above 800 °C
	high cost
	appreciably better stability than other thermocouples
	small basic tolerances

The thermal voltage may be highly nonlinear with temperature; controllers and recorders manufactured by JUMO therefore incorporate linearization tables which can be activated when a thermocouple is connected.

Terminals and compensating cables are often made from materials with different thermal properties than the actual thermocouple. Any temperature gradient at these transition points produces an additional thermal voltage which introduces errors into the measurement. As mentioned above, the cold junction temperature has to be added by the evaluation electronics. If this is measured incorrectly the measurement is falsified to the same extent. With fast-response thermocouples the thermocouple wire is often welded to the protection tube. This may result in circuit loops to electrical loads or other thermocouples. A remedy consists of providing insulation in the supply to the amplifier and between the measurement inputs.

The resistance of insulating ceramics decreases rapidly above 1000 °C. In spite of the insulated construction, the supply voltage can then leak into the measurement line. Contamination of a thermocouple alters its thermal properties. It introduces permanent errors into the measurement and reduces the lifetime.

Comparison of Resistance Thermometers and Thermocouples. Table 2 lists the main distinctions between resistance thermometers and thermocouples.

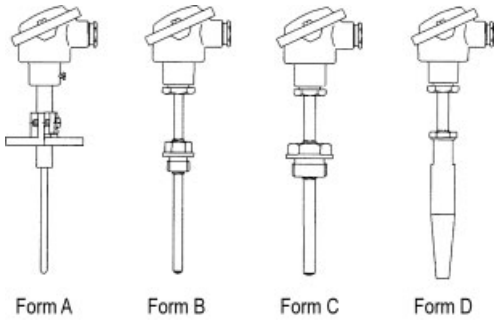

Figure 7. Thermometer fittings for resistance thermometers and thermocouples to DIN 43770 Specification

Construction of Electrical Thermometers. In addition to the almost innumerable number of special versions there are a number of types in which all components are completely described by DIN specifications (Fig. 7). Such a thermometer is of modular construction; it consists of the measurement resistor or thermocouple, the protection tube, the terminal head and the internal terminal base, as well as possibly flanges, clamping devices and movable sleeves. The actual temperature transducer or sensor is that part of the thermometer which is directly

Table 2. Comparison of resistance thermometers and thermocouples

Criterion	Thermocouple	Resistance thermomoeter
Range	− 270 to + 2800 °C	− 220 to + 850 (1000) °C
Overall system accuracy	± 0.75 % of measured temperature	± 0.5 % of measured temperature
Stability: aging error per annum	a few °C	± 0.5 °C
Insulation resistance	important (insulated construction)	very important (large effect)
Size (without protection tube)	relatively small size	relatively large size
Cold junction temperature	required	not required
Characteristic	often strongly nonlinear	slightly nonlinear
Manufacture	relatively simple	relatively complicated

affected by the temperature and which contains the temperature-sensitive part of the thermometer.

General Aspects of Temperature Measurement with Electrical Thermometers. One of the most important errors in temperature measurement with electrical thermometers is a consequence of the thermal inertia of the sensor. If there is a change in the temperature of the medium, after a certain finite time the temperature of the medium and that of the thermometer equalize as a result of the flow of heat.

If the thermometer is considered to be a homogeneous mass m with very good thermal conductivity and surface area A, this can be represented by a first-order curve. The response $\vartheta(t)$ of such a "theoretical thermometer" to a step change $\Delta\vartheta$ is given by the equation:

$$\vartheta(t) = \Delta\vartheta\left(1 - e^{\frac{t}{T}}\right) \tag{4.9}$$

where

$$T = \frac{1}{\alpha}\frac{m}{A}c \tag{4.10}$$

c = specific heat capacity of the sensor material, α = heat-transfer coefficient.

After a step change in the temperature of the medium, the subsequent temperature changes of the sensor take the form of a pure exponential function, the shape of the curve being determined only by the time constant T. From Equation (4.10) for the time constant, certain properties necessary for a rapid response of the sensor can be deduced. A small time constant (and thus a rapid response of the sensor) results from:

- Good heat transfer between the medium and the sensor
- Low mass of the sensor
- Large heat transfer surface area
- Small heat capacity of the sensor

When measuring with electrical thermometers, it is often not realized that the heat transfer coefficient of a gaseous medium is considerably lower than that of a liquid. Therefore, the response time of a thermometer used for measuring the temperature of gases is considerably longer than that for liquids.

In practice, electrical thermometers give curves of higher order. This is especially the case if the sensor element is inside a protective tube. The higher order is the result of the series arrangement of the many individual elements (steel, air, silica sand, sensor) that are separated by thermal resistances.

To compare electrical thermometers with respect to their dynamics, manufacturers determine the response to a step change under defined conditions (medium, flow rate, etc.). The dynamic characteristics are then determined from the delay curve of the measured temperature. The times are usually determined at which the measured temperature rises to 50 % ($t_{0.5}$) and 90 % ($t_{0.9}$) of the step change in the temperature of the medium. These response times are usually given for various media (water, oil, etc.).

It is acceptable to regard a thermometer as having a first-order curve in order to give an approximation of its behavior. However, as already mentioned, electrical thermometers have curves of higher order, so that a time constant cannot be determined and is therefore not quoted by manufacturers.

A thermometer is rarely used for the measurement of ambient temperature. The measurement point is generally warmer or colder than the environment. It is then possible for some heat to be conducted through the thermometer, leading to incorrect measurements. The sensor is connected through cables which are good con-

ductors of heat and which form a thermal link. This problem can be overcome by good thermal connection to the protection tube while at the same time decoupling the connection cables. The probe housing should incorporate as much insulation as possible. The immersion depth of the thermometer in the fluid as well as the flow velocity of the fluid should be chosen to be adequately large.

The Radiation Pyrometer. Every surface whose temperature is above absolute zero emits electromagnetic radiation (thermal radiation) in the wavelength range $0.1 < \lambda < 1000\,\mu\text{m}$. Measuring devices that can determine the temperature of a body from its thermal radiation are known as pyrometers.

If a radiation flux impinges on a surface, it is partly reflected, partly absorbed, and partly transmitted. The relevant proportions are defined as follows:

$$\text{Reflectance} \quad r = \Phi_\text{r}/\Phi \leq 1 \tag{4.11}$$

$$\text{Absorbtance} \quad \alpha = \Phi_\text{a}/\Phi \leq 1 \tag{4.12}$$

$$\text{Transmittance} \quad \tau = \Phi_\text{tr}/\Phi \leq 1 \tag{4.13}$$

The magnitude of these fractions depends very much on the bodies being irradiated. For a so-called black body, all the radiation is absorbed and converted into sensible heat. However, a body does not only absorb radiation, but also emits it because of its own temperature. The black body or the black emitter has the highest capacity to emit. The ratio of the radiation emitted by a nonblack body to that emitted by a black body is termed the emissivity. The emissivity depends on the surface temperature and the wavelength of the radiation. According to Kirchhoff, the emissivity is the same as the absorptance α. Bodies for which the emissivity is independent of the temperature and the wavelength are known as gray emitters. However, real bodies should normally only be termed gray emitters over a limited spectral range.

The radiation flux is defined as the specific-spectral emission Φ of the black body relative to wavelength and surface area:

$$M_{\lambda b} = \left[\frac{\text{W (power)}}{\text{m}^2\,\text{(surface area)} \cdot \text{m (wavelength)}}\right] \tag{4.14}$$

According to Planck's Law, the specific-spectral emission $M_{\lambda\text{b}}$ of a black body is a function of temperature and wavelength (see Fig. 8).

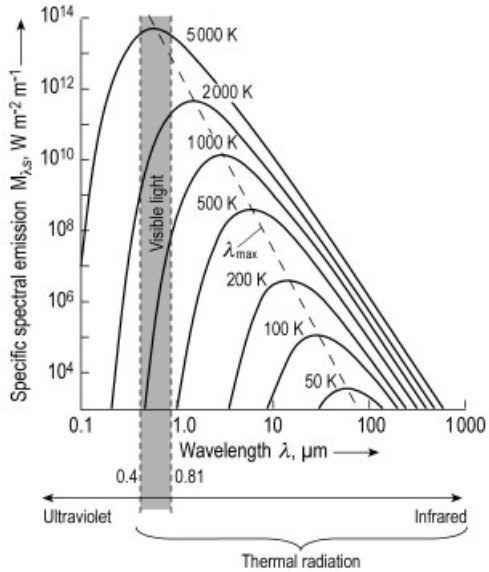

Figure 8. Specific spectral emission of a black body as a function of wavelength

However, no body emits as an ideal black body, and the emission of a nonblack body is therefore less than that of a black body. The emission of a body is given by:

$$M_\lambda = \varepsilon_\lambda M_{\lambda s} \tag{4.15}$$

where ε_λ is the spectral emissivity.

If the spectral emissivity of a body is known and if one considers the radiation in a narrow wavelength band of the spectrum, the temperature of the body can be found from its radiation and from Planck's Law.

It is often desired to determine the surface temperature not from the radiation in a particular spectrum but from the sum of the total radiation (all wavelengths). The Stefan – Boltzmann Law leads to the following equation:

$$M_s = \varepsilon C_s \left(\frac{T_\text{K}}{100}\right)^4 \tag{4.16}$$

where T_K is the temperature in Kelvin and C_S a proportionality constant.

The above physical law is utilized for remote temperature measurement.

5. References

General References

1. T. D. McGee: *Principles and methods of temperature Measurement,* John Wiley & Sons, New York 1988.
2. F. Bernhard: *Handbuch Technische Temperaturmessung,* Springer, Heidelberg 1998.
3. For the practician: Elektrische Temperaturmessung, JUMO (can only be ordered directly from the company JUMO, Fax. No. ++49(0)661/6003-604)

Tequila → Spirits

Cork

Luis Gil, Instituto Nacional de Engenharia e Tecnologia Industrial, Lisboa, Portugal

Cristina Moiteiro, Instituto Nacional de Engenharia e Tecnologia Industrial, Lisboa, Portugal

1. Introduction 2
1.1. Definition and Origin 2
1.2. History 2
1.3. Morphology 3
1.3.1. Microscopic Aspects 3
1.3.2. Macroscopic Aspects 4
1.4. Chemical Composition 4
1.4.1. Suberin 4
1.4.2. Lignin 6
1.4.3. Polysaccharides 6
1.4.4. Waxes 6
1.4.5. Tannins 7
1.4.6. Other Components 7
1.5. Physical and Mechanical Properties 7
2. Cork Extraction 11
3. Production of Cork Based Products 11
3.1. Disks and Cork Stoppers 11
3.2. Granulates and Broken 13
3.3. Composition Cork 13
3.3.1. Floor and Wall Coverings 13
3.3.2. Agglomerated Disks and Cork Stoppers 13
3.3.3. Corkrubber 13
3.4. Insulation Corkboard 14
3.5. Other Specific Products 14
4. Uses 14
4.1. Stoppers 14
4.2. Civil Construction 15
4.3. Industrial Applications 15
4.4. Automotive Industry 15
4.5. Other Specific Uses 15
5. Quality and Standardization 15
5.1. Standardization and Testing 15
5.2. Quality in Cork Production 16
6. Economic Aspects 16
6.1. Forestall Production 16
6.2. Personnel and Companies 16
6.3. Consumption of Products and Markets 17
6.4. Producers 17
7. Environmental and Toxicological Aspects 17
7.1. Industrial Wastes 17
7.2. Recycling and Treatments 18
7.3. Emissions 18
8. Legal Aspects 18
9. References 18

1. Introduction

1.1. Definition and Origin

Cork is the suberous covering (suberose parenchyma, or bark) of the species *Quercus Suber* L., commonly known as the cork oak. Cork is composed of an aggregate of cells, about 42 millions per cubic centimeter, which have five layers walls. Cork is a very light material, elastic and flexible, and impermeable to gases or liquids. It is imperishable and a good electric, thermal, sound, and vibration insulator [1].

The cork oak grows mainly in the Mediterranean zone, especially where there is an Atlantic influence, and Portugal produces more than half of the world's cork. The cork tree has a natural reproduction by acorns or, more frequently by producing shoots [2]. The cork oak is a dicotyledon of the family Fagaceae, and it is part of the genus Quercus having dozens of varieties [2].

All the trees have a thin layer of cork in their bark but *Quercus suber* is the only one where the cork forms a several centimeters thick layer around the trunk, with the function of insulating the tree from heat and loss of moisture and protecting from damage by animals and fire. Cork regenerates after each stripping.

Today cork products are used for thermal insulation in refrigerators and rockets, acoustic insulation in submarines and recording studios, seals and joints in woodwind instruments and combustion engines, as energy-absorbing medium in flooring, shoes, and packaging, and of course as stoppers [3]. The utilization and processing of commercial cork depend very much on its structural quality [4].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.f07_f01

1.2. History

The cork oak dates from the Oligocene stage of the Tertiary Period. It is thought that it was disseminated from the region covered by the Tyrrhenian Sea and that it migrated to the Iberian region in the Miocene stage [2].

Strangely, the first peoples to put cork to what it became after were not of these regions. In ancient China (3000 B.C.) cork was used for fishing buoys, and amphoras were found closed with cork in excavations in Pompeii and in tombs in Egypt [5]. In ancient Greece the bark of the cork was used to make buoys for fishing nets, sandals, and stoppers for wine and olive oil vessels.

In the early days of Portuguese nation, cork was used in building (including shipbuilding). By the 14th century it was already an economic asset of national importance in Portugal. The great step forward for the generalized use of cork was taken by the French Benedictine monk Dom Pérignon (1639 – 1715), proctor of the Hautvillers' Abbey who developed champagne production. He observed that the wooden stoppers wrapped and soaked in olive oil, that were used for the containers holding sparkling wine often jumped out. So he used cork stoppers with surprisingly good results and soon cork became essential for wine producers [2].

In 1750 in Angullane (Gerona), close to the French border, the first stopper factory opened, marking the beginning of cork industry, which then expanded to Portugal. Around 1770, when Port Wine trade had begun to flourish, the cork stopper became indispensable.

The real boom of the cork industry began in the last quarter of the 19th century. In 1915 Portuguese cork oak forests covered between 300 000 and 470 000 ha and nowadays encompass more than 700 000 ha. In 1927 Portuguese annual cork production was 16 000 t and is now about ten times higher [2].

At the end of the 19th century, an American manufacturer of lifejackets accidentally found that it was possible to produce agglomerated cork. This was the beginning of a new world of opportunities in which waste cork and cork previously thought to be of no commercial value could also be used [6].

In the first half of the 20th century, the world cork market was controlled by a handful of multinational companies. However, their productive investment was minimal, especially in Portugal and Spain. They worked in these countries through small offices that acted as purchasing centers for the raw materials and some manufactured goods, that were channeled to other centers for processing and distribution. In the 1960s, Portuguese companies started to invest in modern factories and today Portuguese cork industry processes more than 2/3 of world's cork [1]. Nowadays Spain is the second world cork producer and transformer.

1.3. Morphology

1.3.1. Microscopic Aspects

Cork has an open structure, which is made up of cells of 20 to 30 millionths of a millimeter in size, forming the suber and the rays. The cell lumen contains mainly nitrogen. A thickness of 1 mm corresponds to about thirty layers of cells, 1 mm^3 has about 20 000 – 40 000 cells [2,7].

Cork tissue is made up of dead micro-cells, generally 14-sided polyhedrons, slotting in one against the other. The contents of newly formed cells disappear during growth and the subsequent suberization process of the membranes, which imparts the completion of communication with the plant living tissues.

The cells are staked in long rows with thin walls. The radial section also shows five, seven and eight sided shapes [3]. Cells walls are covered with very thin layers of waxes and suberin which make them impervious to air and water and resistant to attack by many acids. The cell walls are conjugated. The cork cells are closed hexagonal prisms (average) stacked in rows so that two cells share the hexagonal faces, and the rows are staggered so that the membranes forming the hexagonal faces are not continuous across rows (see Fig. 1).

The cellular structure of cork is significantly influenced by temperature; upon heating cork cells expand and the cell walls stretch, attaining their maximum volume (twice the original volume) at 250 °C [8].

The intracellular layers comprise five parts: two made of cellulose, a further two of hard matter with impermeable components (suberin, waxes) and a final ligneous part, whose function is to maintain structure and rigidity [1].

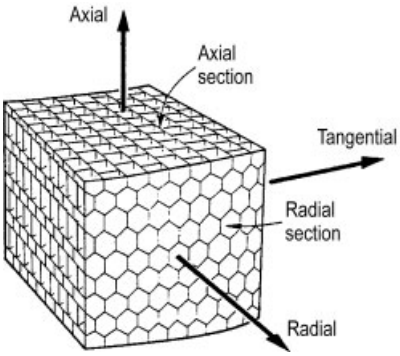

Figure 1. Diagram of cork tree and cork with axis system

The cell walls have a thickness of about 1 – 2 µm, a height of about 10 – 45 µm and a hexagonal face edge of about 20 µm (see Fig. 2). The density of cork is related to that of the cell wall material, which is close to 1150 kg/m^3 [3,7]. The thickness of cork cells is lower in spring and greater in autumn. The cell height has also a minimum value in autumn.

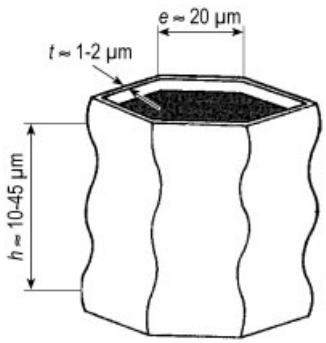

Figure 2. Micrograph of cork cells and corrugated cell showing dimensions

The unevenness of membrane thickness, height and diameter of the cells can affect the cork's mechanical and physical properties, namely compressibility and elasticity. The prismatic shape of the cells makes the cork anisotropic [7].

When cork deforms, cell walls bend and buckle, giving resilience. Tensile deformation along the prism axis unfolds the corrugations, straightening prism walls. Compressive deformation folds the corrugations up to a certain point where the cells collapse completely. Axial compression produces almost no lateral expansion, due to the fact that cells fold [3,7].

1.3.2. Macroscopic Aspects

The cork tree has a remarkable capacity to create a suberous tissue from its inner bark. The external productive layer or phloem grows in the direction of the bark and produces cork [6]. As the phellogen is only active for 6 – 7 months each year, the differences in cell size and the thickness of cell membranes between cork produced in different year seasons leave discernible rings (growing rings), which change from 1.5 to 4 mm in thickness on average, mainly according to the nature of the soil, climatic conditions, and vitality of the tree [6]. Younger trees produce usually very porous and cracked cork and old and aging trees form thin cork.

Natural cork contains lenticels, which are tubular channels that connect the outer surface of the bark to its inner surface allowing oxygen into and carbon dioxide out from the new cells that grow there.

The life cycle of the cork oak produces three qualities of suberose tissue: virgin cork, reproduction cork from the second stripping, and reproduction cork from subsequent stripping. *Virgin cork* is the suberose tissue extracted for the first time (usually when the tree is 25 years old), it is hard and has an irregular structure. *Summer virgin* is the cork extracted from a level above the previous extraction. *Winter virgin* is the virgin cork from the pruned branches, containing fragments of fiber and ligneous tissue. The first *reproduction cork* (taken at least nine years later) is more regular than virgin cork but does not yet have quality enough for cork stopper production [1]. Only the second reproduction cork can be used for cork stoppers. All the other types of cork can be used for agglomerates.

The thickest suberose layer is generally formed in the growing cycle subsequent to cork extraction and then the cork produced diminishes progressively year after year.

Porosity is a consequence of the lenticels, orifices or pores in the tissue. Porosity depends on the area occupied by lenticular channels cut tangentially. If porosity is excessive (>4 %) the economic value of cork may decrease for some applications [1].

1.4. Chemical Composition

The chemical constitution of cork depends on factors such as geographical origin, climate and soil conditions, genetic origin, tree dimension, age and growth conditions [1,9,10].

According to the proposed model [11], the cellular structure walls of cork consists of a thin and lignin-rich middle-lamella primary wall, a thick secondary wall made up from alternating suberin and waxes lamellae and a thin tertiary wall of lignin and polysaccharides. Studies on chemical composition show that the secondary wall is lignified and therefore may not consist exclusively of suberin and waxes [9]. Further information on suberin structure has been obtained from solid-state ^{13}C NMR spectroscopy studies, showing that the aliphatic chain of suberin is spatially separated from lignin and polysaccharides [12].

The main chemical constituents of cork can be divided in five groups:

Suberin	33 – 50 %
Lignin	12.6 – 30 %
Cellulose and polysaccharides	12 – 20 %
Waxes	3.5 – 7.9 %
Tannins and other phenolic substances	6 – 7 %

1.4.1. Suberin

The structure of suberin in cork is not yet fully understood. It has been proposed that suberin consist of a polyester structure composed of long-chain fatty acids, hydroxyfatty acids and phenolic moieties, linked by ester groups, as shown in Figure 3 [13].

Suberin is insoluble in all solvents but can be depolymerized by cleavage of the ester linkages by alkaline hydrolysis, alcoholysis, transesterification (catalyzed by soda, alkoxides or boron trifluoride) or reduction. The aliphatic components of suberin include fatty acids, fatty alcohols (mainly C_{20}, C_{22}, and C_{24} alcohols) and glycerol. Glycerol (**1**), is present as component of esters of suberinic acids [14–18].

The fatty acids include monoacids, α,ω-diacids and ω-hydroxyacids, which can have different functionalities including double bonds, vicinal hydroxy and epoxide groups. The most

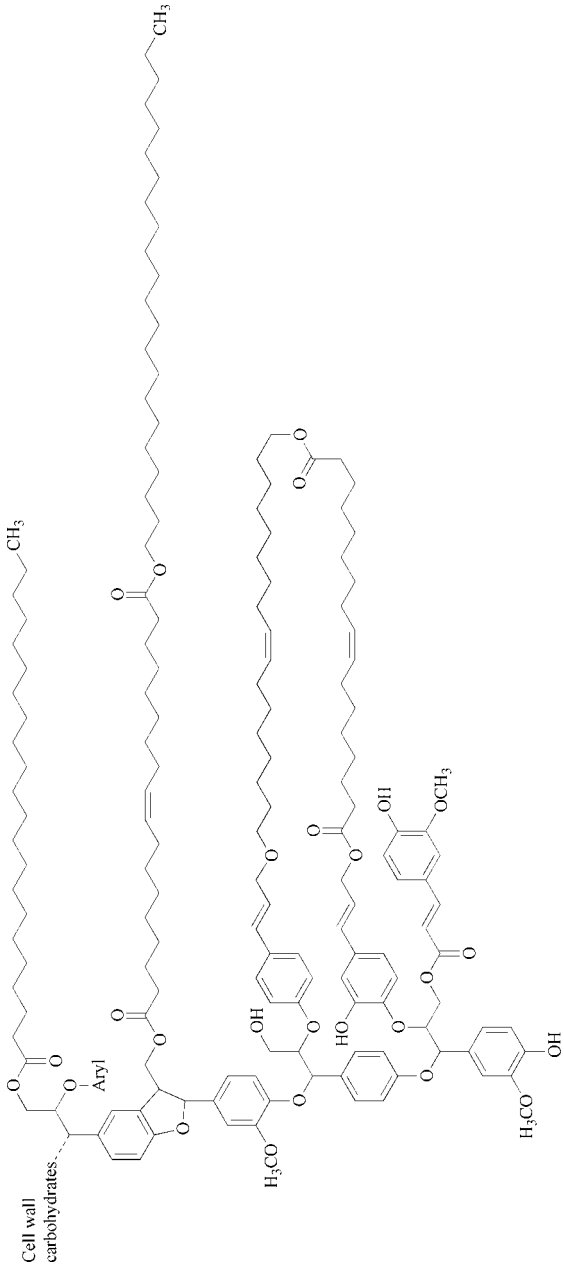

Figure 3. (rotated) Schematical structure of cork suberin

important acids are 9,10-dihydroxyoctadecanedioic (phloionic acid) (**2**), 18-hydroxyoctadec-9-enoic (**3**), octadec-9-enedioic (**4**), docosanedioic (phellogenic acid) (**5**), 22-hydroxydocosanoic (phellonic acid) (**6**), and 9,10,18-tridihydroxyoctadecanoic acids (phloionolic acid)

Table 1. Composition of suberin monomers of cork (%)

Compounds	Structure	Content in cork, %
Aliphatic		
Propane-1,2,3-triol (glycerol) **1** [*56-81-5*]	$HO\text{-}CH_2\text{-}CH\text{-}OH\text{-}CH_2OH$	2.0 – 7.0
1-Alkanos (C_{18} – C_{28})	$CH_3(CH_2)_nOH$	2.3 – 2.8
Alkanoic acid (C_{16} – C_{28})	$CH_3(CH_2)_nCOOH$	1.3 – 10.0
α,ω-Alkanodioic acids (C_{16} – C_{26})	$HOOC(CH_2)_nCOOH$	6.2 – 12.3
α,ω-Alkenodioic acids (C_{18} : 1 – C_{22} : 1)	$HOOC(CH_2)_nCH{=}CH(CH_2)_mCOOH$	2.8 – 25.4
ω-Hydroxyalkanoic acids (C_{16} – C_{26})	$HOCH_2(CH_2)_nCOOH$	10.8 – 33.0
ω-Hydroxyalkenoic acids (C_{18} : 1 – C_{22} : 1)	$HOCH_2(CH_2)_nCH{=}CH(CH_2)_mCOOH$	11.4 – 25.7
C_{16} Functionality acids		
7,16-Dihydroxyhexadecanoic acid **8**	$HOCH_2(CH_2)_8CHOH(CH_2)_5COOH$	2.1 – 4.0
8,16-Dihydroxyhexadecanoic acid **9**	$HOCH_2(CH_2)_7CHOH(CH_2)_6COOH$	
C_{18} Functionality acids		
Octadec-9-en-1,18-dioic acid **10** [*2027-47-6*]	$HOOC(CH_2)_7CH{=}CH(CH_2)_7COOH$	2.8 – 10.2
9,10-Epoxy-octadecane-1,18-dioic acid **11**	$HOOC(CH_2)_7HC\text{—}CH(CH_2)_7COOH$ (O)	4.0 – 15.9
9,10-Epoxy-18-hydroxy-octadecanoic acid **12** [*3233-92-9*]	$HOH_2C(CH_2)_7HC\text{—}CH(CH_2)_7COOH$ (O)	5.0 – 15.3
9,10,18-Trihydroxy-octadecanoic acid (phloionolic acid) **7** [*496-86-6*]	$HOCH_2(CH_2)_7CHOHCHOH(CH_2)_7COOH$	2.0 – 5.4
9,10-Dihydroxy-octadecanedioic acid (phloionic acid) **2**	$HOOC(CH_2)_7CHOHCHOH(CH_2)_7COOH$	2.0 – 8.8
Aromatic		
4-Hydroxy-3-methoxycinnamic acid (ferulic acid) **13** [*537-98-4*]	$HOC_6H_3(OCH_3)CH{=}CHCOOH$	7.0
3,5-Dimethoxy-4-hydroxycinnamic acid **14** (sinapinic acid) [*530-59-6*]	$HOC_6H_2(OCH_3)_2CH{=}CHCOOH$	
4-Hydroxy-2-methoxy-benzoic acid **15**	$4\text{-}(HO)C_6H_3 - 2\text{-}(OCH_3)COOH$	
4-Hydroxy-3-methoxybenzaldehyde (vanillin) **16** [*121-33-5*]	$4\text{-}(HO)C_6H_3 - 3\text{-}(OCH_3)CHO$	
Ferulic acid esterified with methyl fellonate **17**	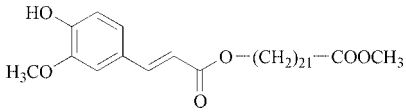	

(**7**). The composition of suberin monomers is shown in Table 1 [14, 15, 19–27].

Several triterpenes and sterols constituents of waxes were also isolated from suberin [19].

The composition of the aromatic fraction of cork suberin is not completely elucidated. The structure of the phenolic part displays features similar to those of lignins. Alkali-depolymerized cork contains large amounts of phenolic compounds (ca. 60 %) [28]. Aromatic acids, not occuring in lignin, have been identified, such as 4-hydroxy-3-methoxycinnamic (ferulic) (**13**), 3,5-dimethoxy-4-hydroxycinnamic (sinapinic) (**14**), and 4-hydroxy-2-methoxybenzoic acids (**15**), together with 4-hydroxy-3-methoxybenzaldehyde (vanillin) (**16**), and esters of the ferulic acid esterified with fellonic (**17**) and other suberinic acid [14, 17, 29–31].

1.4.2. Lignin (→ Liginin)

Although several attempts have been made to extract and characterize cork lignin its structure has not been fully established. The nature of linkages between lignin and aromatic units of suberin is unclear. It is not possible to isolate a "Milled Wood Lignin" (MWL) by the extraction of ball-milled cork with aqueous dioxane at room temperature, applied for the isolation of lignin from wood [32], indicating that, if any "lignin-like polymers" occur in cork, they must be covalently bound to suberin [33], [46]. Furthermore, the calculation of 40 % "lignin" in cork from the methoxyl content (5.6 %) of whole cork [46], does not take into account the methoxy groups in hemicelluloses, e.g., 4-O-methyl-glucuronoxylan, as well as those in suberin itself, e.g., ferulic acid esterified with methyl fellonate (**17**, Table 1).

When ground, extracted reproduction cork was heated for 4 h at 160 °C with ethanol/water (1 : 1, v/v), an "Organosolv Cork Lignin-Like Polymer" (OCL) was obtained in just 2.6 % yield, that still contained "residual aliphatic moieties" [40].

In order to hydrolise the assumed ester bonds between lignin and suberin (Figure 4, see next page), ground reproduction cork was first treated with sodium methylate in methanol and the residual cork (34.4 %) then treated by the Björkman procedure, to yield 4.1 % "Saponified Milled Cork Lignin" ($MCL_{sap.}$), which, based on the results of analytical pyrolysis, was claimed to be similar to softwood MWL [42]. However, a comparison of $MCL_{sap.}$ with MWL from spruce wood by thioacidolysis revealed that $MCL_{sap.}$ contained 34 % less β-O-4 linkages and, by permanganate oxidation, gave 21 – 23 % less oxidation products than the spruce MWL [43].

As, β-O-4 linkages are indicative for the discrimination between lignin and bark polyphenols, taking into account the low yields of permanganate oxidation products as well as those of OCL and $MCL_{sap.}$, it may be well concluded that the aromatic moieties in cork, linked to suberin, are to be classified as bark polyphenols rather than lignins.

1.4.3. Polysaccharides

Cellulose represent ca. half of the total polysaccharides of cork (5 – 10 %). Hemicelluloses i.e., heteropolymers, which include several different sugar moieties (arabinose, xylose, glucose, galactose, etc.) constitute the other half [15].

Three types of hemicelluloses are found in cork: hemicellulose A and two hemicelluloses B (B-1 and B-2). Hemicellulose A contains xylose (95.2 %) and 4-*O*-methylglucuronic acid (5.9 %) [40]. Hemicellulose B-1 is composed of xylose, 4-*O*-methylglucuronic acid, arabinose, galactose, mannose, and glucose in the molar ratios 135 : 12 : 7 : 11 : 2 : 30 [41]. Hemicellulose B-2 is more highly branched than hemicelluloses A and B-1 and contains xylose, arabinose, glucose, galactose, 4-*O*-methylglucuronic acid, and rhamnose in the molar ratio of 17 : 12 : 12 : 6 : 4 : 1 [42]. The model of the hemicellulose structure from cork of *Quercus suber* L. is depicted in Figure 5 (see below Table 2 on pages after next) [43].

1.4.4. Waxes

The neutral compounds, extracted from cork by apolar solvent, include *n*-alkanes ($C_{16} - C_{34}$), fatty alcohols (with even numbers of C-atoms from $C_{18} - C_{26}$, containing traces of intermediate alcohols with an odd number of C-atoms and some unsaturated alcohols), glycerol, triglycerides, triterpenes of the friedelane and lupane families, and a very small fraction of monocarboxylic fatty acids (C_{16} and C_{24}). The triterpenes represent ca. 50 % of waxes and 2 – 5 % of cork [10,15,44].

The main components of the acid fraction are fatty acids, saturated even numbered C_{14} – C_{24}, odd C_{15}, C_{17}, C_{21}, and ω-hydroxyacids, triterpenes, sterols, and C_{20} and C_{24} alcohols. Table 2 (see pages after next) shows the chemical composition of the cork waxes, chemical and common names and structures [9,10,40–52].

1.4.5. Tannins

The tannins are phenolic and polyphenolic compounds which can be extracted from cork by polar solvents, such as water. The phenolic compounds may be monomeric (phenols, benzoic and cinnamic acids) and polymeric (hydrolyzable and condensed tannins). *Hydrolyzable tannins* are composed mainly of trihydroxybenzoic acid (gallic acid) (**56**) [*149-91-7*] or hexahydroxydifenic acid esterified with glucose (**57**). *Condensed tannins* originate from flavonoid type precursors by polycondensation reactions [44].

The phenolic compounds extractable from cork represent ca. 6 – 7 % of total cork corresponding to 95 – 98 % of tannins.

1.4.6. Other Components

The ash content of cork is ca. 1 – 3 %. Calcium is the most abundant inorganic constituent, in addition cork contains potassium, phosphorus and magnesium [1,53].

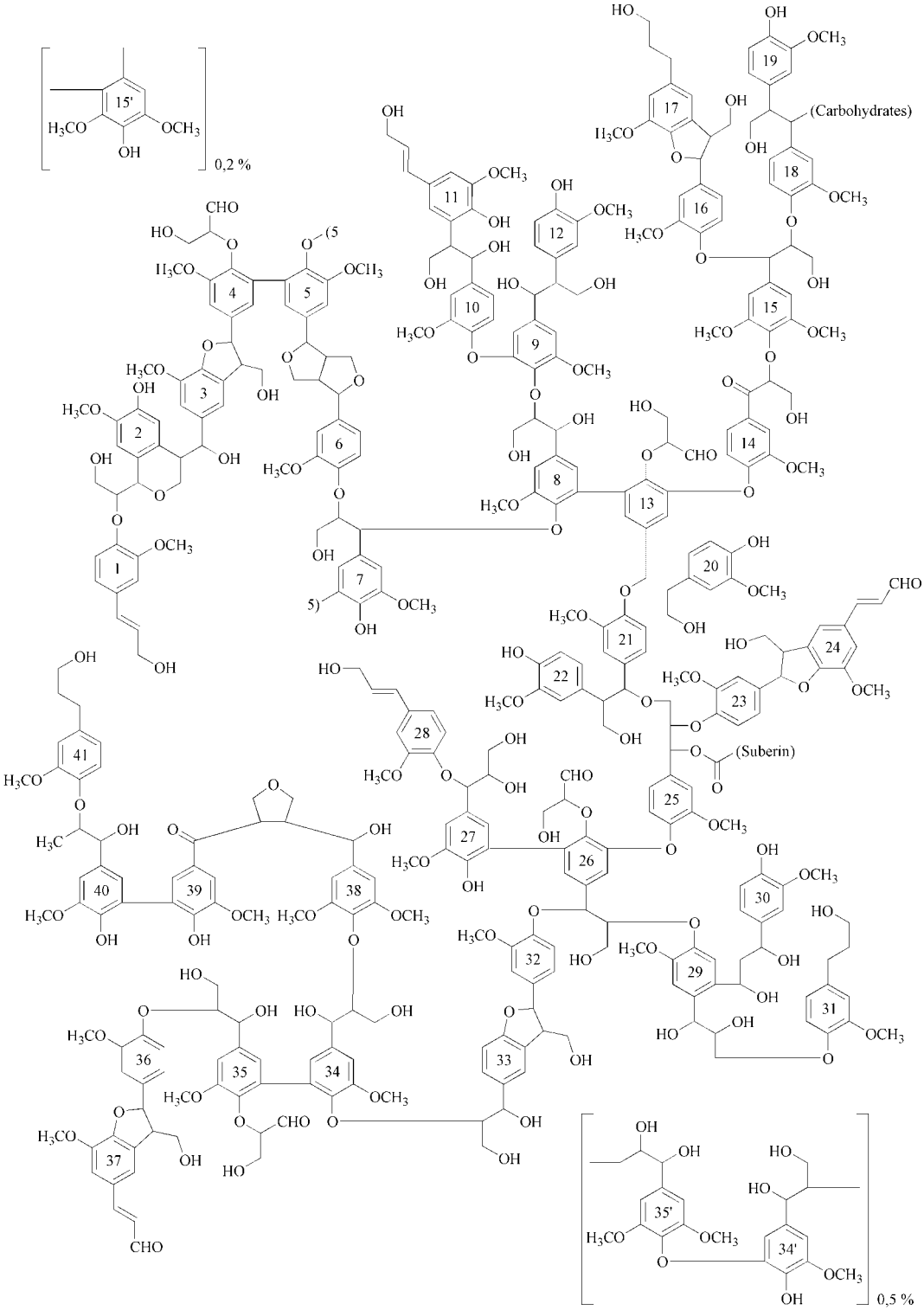

Figure 4. Model of the lignin structure of *Quercus suber*

Table 2. Chemical composition of cork waxes of *Quercus suber*

Compounds	Structure	Content in cork, %
Fatty alcohols	$H_3(CH_2)_nCH_2OH$	0.7
Octadecanol **18** [*112-92-5*]	$n = 16$	
Icosan-1-ol (eicosanol) **19** [*629-96-9*]	18	
Heneicosan-1-ol **20**	19	
Docosan-1-ol **21** [*661-19-8*]	20	
Tricosan-1-ol **22** [*3133-01-5*]	21	
Tetracosan-1-ol **23** [*506-51-4*]	22	
Pentacosan-1-ol **24** [*26040-98-2*]	23	
Hexacosan-1-ol **25** [*506-52-5*]	24	
Propane-1,2,3-triol (glycerol **1** [*56-81-5*]	$OHCH_2CH_2OHCH_2OH$	
Triterpenes	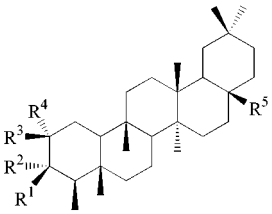	2 – 5
Friedelan-3-one (friedelin) **26** [*559-74-0*]		
3-Hydroxyfriedel-3-en-2-one **27**		
2α-Hydroxyfriedel-3-one (cerin) **28** [*468-67-7*]		
2α-Acetoxyfriedel-3-one (cerin acetate) **29**		
2β-Acetoxyfriedel-3-one **30**		
3α-Hydroxyfriedelan-2-one **31**		
28-Hydroxyfriedelan-3-one (canophyllol) **32**		
	26 $R^1, R^2 = O; R^3, R^4 = H; R^5 = H$	
	28 $R^1, R^2 = O; R^3 = H; R^4 = OH; R^5 = H$	
	29 $R^1, R^2 = O; R^3 = H; R^4 = OAc; R^5 = H$	
	30 $R^1, R^2 = O; R^3 = OAc; R^4 = H; R^5 = H$	
	31 $R^1 = H; R^2 = OH; R^3, R^4 = O; R^5 = H$	
	32 $R^1, R^2 = O; R^3, R^4 = H; R^5 = CH_2OH$	
	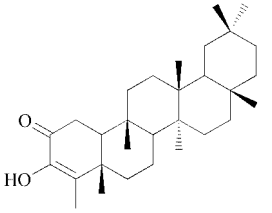	
	27	
Lupane	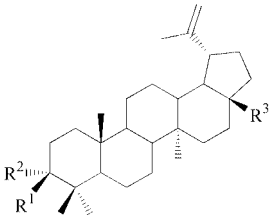	
20(29)-Lupen-3β-28-diol (betulin) **33** [*473-98-3*]		
3β-Hydroxy-20(29)-lupen-28-oic acid (betulinic acid) **34** [*472-15-1*]		
3-Oxo-20(29)-lupen-28-oic acid (betulonic acid) **35** [*4481-62-3*]		
Lup-20(29)-en-28-ol-3β-yl-caffeate **36**		
	33 $R^1 = OH; R^2 = H; R^3 = CH_2OH$	
	34 $R^1 = OH; R^2 = H; R^3 = COOH$	
	35 $R^1, R^2 = O; R^3 = COOH$	
	36 $R^2 = H; R^3 = CH_2OH; R^1 =$	
	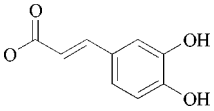	

Table 2. Continued.

Compounds	Structure	Content in cork, %
Sterols β-Sitosterol **37** [*83-46-5*] Campesterol **38** [*474-62-4*]	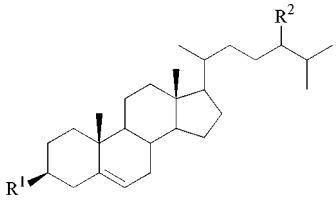 **37** $R^1 = OH$; $R^2 = CH_2CH_3$ **38** $R^1 = OH$; $R^2 = CH_3$	
Suberindiol **39**	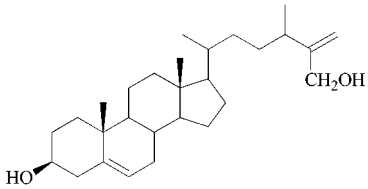 **39**	
Fatty acids	$H_3C(CH_2)_nCOOH$	1.1
Tetradecanoic acid **40** [*544-63-8*]	$n = 12$	
Pentadecanoic acid **41** [*1002-84-2*]	13	
Hexadecanoic acid **42** [*57-10-3*]	14	
Heptadecanoic acid **43** [*506-12-7*]	15	
Octadecanoic acid **44** [*57-11-4*]	16	
Icosanoic acid (Eicosanoic acid) **45** [*506-30-9*]	18	
Heneicosanoic acid **46** [*2363-71-5*]	19	
Docosanoic acid **47** [*112-85-6*]	20	
Tetracosanoic acid **48** [*557-59-5*]	22	
Heptadec-9-enoic acid **49** [*10136-52-4*]	$H_3C(CH_2)_6CH{=}CH(CH_2)_7COOH$	
Hexadec-12-enoic acid **50**	$H_3C(CH_2)_2CH{=}CH(CH_2)_{10}COOH$	
Heptadec-12-enoic acid **51**	$H_3C(CH_2)_3CH{=}CH(CH_2)_{10}COOH$	
Octadec-9-enoic acid [*2027-47-6*] **52**	$H_3C(CH_2)_7CH{=}CH(CH_2)_7COOH$	
Octadeca-9,12-dienoic acid [*2197-37-7*] **53**	$H_3C(CH_2)_4CH{=}CHCH_2CH{=}CH(CH_2)_7COOH$	
18-Hydroxy-octadeca-9,12-dienoic acid **54**	$HOCH_2(CH_2)_4CH{=}CHCH_2CH{=}CH(CH_2)_7COOH$	
18-Hydroxy-octadec-9-enoic acid **3**	$HOCH_2(CH_2)_7CH{=}CH(CH_2)_7COOH$	
***n*-Alkanes** 2,6,10,15,19,23-Hexamethyl-2,6,10,14,18,22-tetracosane-hexaene (squalene) [*111-02-4*] **55**	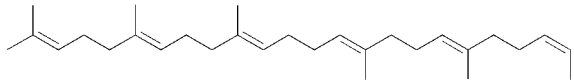	
$C_{20} - C_{34}$	$CH_3(CH_2)_nCH_3$	0.6

1.5. Physical and Mechanical Properties

Cork has a low density, which usually varies from 120 – 140 kg/m^3 depending on the region, the quality of the tree and the degree of humidity. Another special characteristic of cork is its great elasticity. When cork is rapidly compressed under elastic conditions it returns immediately to about 85 % of its initial volume and in 24 h about 95 % of its initial dimension are reached.

The average properties of cork are listed below [7]:

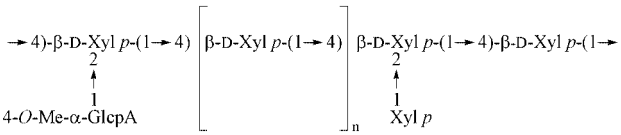

Figure 5. Model of hemicellulose structure of *Quercus suber*

Density	120 – 140 kg/m^3
Young's modulus	4 – 60 MN/m^3
Shear modulus	2 – 30 MN/m^2
Poisson ratio	0.15 – 0.22
Collapse strength (compression)	0.5 – 2.5 MN/m^2
Fracture strength (tension)	0.8 – 3.0 MN/m^2
Loss coefficient	0.1 – 0.3
Coefficient of friction	0.2 – 0.4
Thermal conductivity	0.025 – 0.028 J m K^{-1}

Because of its small Poisson coefficient (approx. 0) compression of cork parallel to the prism axis produces little lateral spreading [7], a property useful in gaskets. Cork is almost impermeable to gases and liquids, mainly due to the hydrophobic nature of the cell membranes. Cork sticks well to smooth surfaces even when humid, due to its elasticity, to the suction effect produced by the broken cells in contact with the walls and traces of resinous substances, aspects which are very important in a stopper. The friction coefficient $\mu_{\text{cork}--\text{glass}}$ is around 0.6.

Figure 6 shows a complete stress – strain curve for cork, depicting all the characteristics of a cellular solid. Cork is linear-elastic in the first stage, then its structure collapses and gives an almost horizontal plateau until the complete collapse of the cells causes the curve to rise steeply.

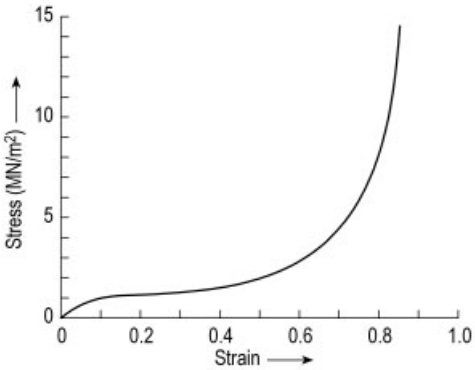

Figure 6. Stress – strain curve for cork

When a sharp object like a pin is stuck into cork, the deformation is highly localized. A layer of cork cells, occupying a thickness of about a quarter of the diameter of the indenter, collapses as it suffers a large strain. The indenter volume is absorbed by the collapse of the cells.

The presence of suberin renders the cork impermeable, and because of its content of tannins cork has no tendency to rot. A high concentration of oxygen is needed to incinerate cork and when the combustion starts (only at very high temperatures) cork burns on the surface and a layer of carbon forms and prevents the fire to spread. One can say it is a self-extinguishing material.

2. Cork Extraction

The first cork extraction (virgin cork) is done after the trunk of a cork oak has reached 0.7 m of perimeter [1].

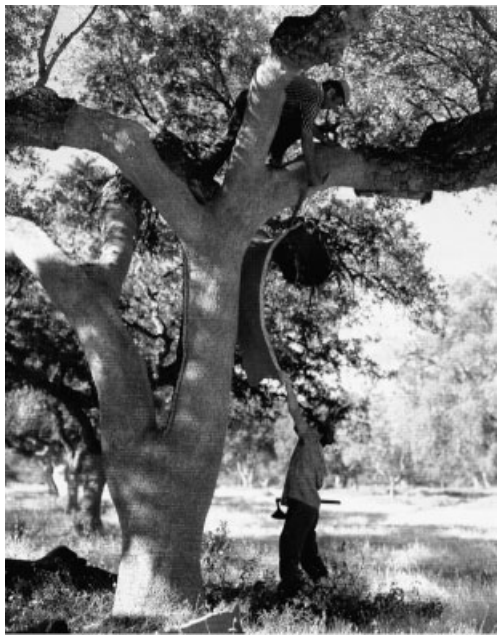

Figure 7. Cork tree and cork extraction

The stripping when done well does not harm the trees because the first layer of reproduction cork merges with the continuously developing virgin layer in the unpeeled part of the tree. The "mother" (trunk surface after peeling) changes from a pink color to red ochre and afterwards to a reddish brown and finally to a gray crust-like formation [2].

After extraction, the reproduction cork is produced at a rate of 1.5 to 4 mm per year. In Portugal it is required by law to wait at least for 9 years after stripping before a new extraction in order to get a useful harvest (about 2 to 5 cm). This thickness is needed because it is related

with stopper diameter. In other locations a cycle of 12 years or even more may be needed. The collection of cork may be repeated every 9 years up to 12 or 15 harvests. Afterward, the tree will no longer be acceptably productive in general. Although yield varies greatly, an adult cork oak, with normal development, can produce from each stripping between 40 to 60 kg of cork [1], although much higher values (> 1 t per strip) can be achieved.

Harvest is between beginning of June and end of August. The cork is extracted in spring or summer because at this time the cork comes away easily from the trunk since the tree is growing and new cork cells are being generated. It is in the summer that the phelogen is in full production of new cork cells and it is easier to extract the bark by the disruption of these new-formed cells. The cut must be carefully done so as not to affect the mother layer. Care must be taken not to damage the cambium, which unlike the phellogen cannot regenerate itself. When wounded it heals very slowly and wood may start to rot. Even when this is avoided a scar will be seen in future cork and reduce its value. In the first harvest the good practice is not to cut higher than two times the perimeter of the tree at chest level and this height may increase in the following peelings up to three times this perimeter. A special axe is used for the harvest. The blade is used to make the incision while the end of the handle is shaped specially to detach the cork [1]. Besides this manual stripping also a mechanical process exists, based on small machines, that work using a blade with a sawing motion.

After the harvest "planks" are obtained that are dried in big-formed piles in the forest for approximately 3 – 6 months [1].

3. Production of Cork Based Products

3.1. Disks and Cork Stoppers

After harvest and stabilization cork is taken to the factory where the first operation of boiling takes place [1]. The major structural alteration due to boiling is the attenuation of the cork cells corrugation and an expansion of 10 – 15 % in the radial direction and 5 – 7 % in the other directions [54]. Boiling also decreases porosity by about 50 % [55] (see Fig. 8).

In boiling the cork is immersed in boiling water for approximately 1 h, usually in big open tanks with direct fire underneath. The aim of this operation is to remove dirt, insects, etc. to extract contaminating products and to render the planks softer and more flexible, and thus to facilitate the workability. The boiling water is regularly changed (at least every week) in modern systems.

After boiling to mature cork, it is left to ripen for two to four weeks in cellars (or some days after the new boiling process), sometimes under controlled hygrometric conditions. Very few thermophilic species of microflora, which eventually survived boiling and further recontamination, may appear on the planks. Finally, the edges are cut to make the planks rectangular, and the planks are inspected and selected according to thickness and quality (i.e., flexibility, homogeneity, color, evenness, and natural defects). 40 – 50 % of the original cork weight is lost in this step of cork preparation (separation of defects, drying, edge cutting) [1]. The best cork (about 25 %) goes to stopper production and the remainder is used for agglomerates.

The rectangular planks are cut into strips with a thickness corresponding to the diameter of the cork stopper and a width corresponding to the height of the stopper. The corks are punched with a cutting tube (rotating), the machines being more or less automated [1]. As lenticels lie parallel to the cell prism axis, cork cut parallel to this axis will leak. That is why natural cork stoppers are cut with the prism axis (and lenticels) perpendicular to the cork stopper axis [3,56]. After this, dimensional rectification, such as ends trimming and control of height and diameter, is performed.

The corks are classified in four to six categories, this selection being made by hand by specially qualified workers or by classifying machines (electronic optical reading). A good cork should be light and even in color, the pores being light and few in number. Harder corks may be chosen for wine under pressure [1].

The selection of cork stoppers is one of the most important operations in cork processing. Several parameters are analyzed, and a definition of cork defects for stoppers may be as follows [1]:

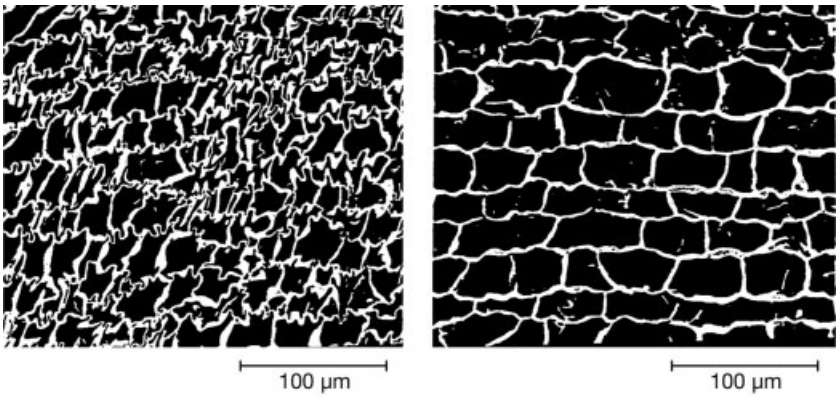

Figure 8. Effect of boiling on the cellular structure of cork

- Belly spots – surface depression caused by the dense inner surface of the cork strip;
- Chips – pieces missing from the cork surface (near the ends) due to dry or brittle cork material or faulty corker jaws;
- Cracks – fissures in the cork surface, which can cause chips;
- Dry year – narrow woody growth rings caused by lack of rain;
- Greenwood – undulation in the cork surface caused by moisture;
- Hardwood – hard rough areas on the cork surface caused by boring too close to the bark surface;
- Porosity – high-density pore activity;
- Worm holes – holes caused by worms.

To improve the appearance, the pores or lenticels can optionally be filled in with a mixture of cork powder and glue. Afterward the corks are washed (in tanks or automatic rotating machines), nowadays normally with a peroxide solution. After washing the corks are dried to 6–8 % moisture content. They can also be satin finished, i.e., covered by a fine film of paraffin (solid, oil emulsion) or silicone to facilitate cork stopper extraction and are marked (ink or fire) [1,56]. All the products used in finishing are of food grade.

Cork stoppers can be sold as finished or semi-finished products. In the first case marked and treated corks are packed in plastic bags under vacuum and SO_2. In the second case raffia or jute bags are usually used.

To produce *cork disks*, the thinnest good quality cork (not thick enough for cork stopper manufacturing) is used. From these planks the belly and the rasp are cut to obtain a clean slab of cork from which disks are cut and then dimensionally finished.

3.2. Granulates and Broken

Waste cork from the stopper industry, low quality cork (refuse), and possibly virgin cork are used to produce *cork granulates*. The granulates are separated and classified according to density and grain size. These cork granulates can be used as a final product in several applications or used as raw material for the production of composition cork. The finest are used to produce linoleum.

Broken is the name of small pieces of cork obtained from the breaking and first trituration of cork waste in order to reduce volume, e.g., for transportation.

3.3. Composition Cork

Composition cork is made of granules, which have been joined together using different synthetic or natural binding agents [1]. The agglutination is made, e.g., with asphalt, rubber, casein, cement, gypsum, plastics, urethane, melaminic, or phenolic resins [2]. Composition cork is usually produced in sheets, rolls, or blocks, or even extruded or molded.

3.3.1. Floor and Wall Coverings (→ Floor Coverings)

Floor and wall coverings made of cork are usually produced in the form of tiles, with different thickness, densities and finishing: simply polished, waxed, varnished or with a poly(vinyl chloride) layer. This last group may use a decorative sheet between the PVC and the agglomerate underneath [1]. A cork layer associated to a medium-density fiberboard (MDF) basis, for example, constitute a new type of flooring, known as floating floor coverings. The finest cork granules are used to produce linoleum [2].

Wall coverings have a density of 200–300 kg/m^3 and floor coverings of 400–500 kg/m^3. A mixture of cork granules and glue and/or other additives is filled into a mold which is then closed and heated usually at more than 120 °C for some hours to produce a block which upon cooling (or not) is then sliced into sheets, which are then dimensionally finished. These sheets can be painted, varnished, or covered with layers of other materials.

3.3.2. Agglomerated Disks and Cork Stoppers

For the production of agglomerated cork stoppers cork waste from the cutting stage (until 90 % of the granulated material) is used or material rejected at the stopper sorting stage. Contaminated material should be avoided.

There are simple agglomerated corks and two types of composite stoppers, the stopper for champagne and sparkling wine (head in agglomerate and two or more disks in the bottom) and the "1 + 1" stopper for the other wines (body in agglomerate and one disk in each end).

Champagne corks comprise an upper part of agglomerated cork made be molding, extrusion, or block cutting and usually two or three disks of natural cork, which are in contact with the wine. This special type of corks has thre distinct parts: the "head" that protrudes out of the bottle, the "wheel" which is the part in the neck of the bottle, and the "mirror" which is the part in contact with the wine [6].

3.3.3. Corkrubber

The production of corkrubber is similar to the production of rubber-like products. Rubber and cork granules are mixed in rolls and the mats obtained are introduced into a mold, which is heated for polymerization. The heating process may take from several hours (in a typical oven) to some minutes (in microwave systems).

The most common corkrubber materials include a portion of cork granulates (60–70 kg/m^3), ranging from 15–260 parts of cork per 100 parts of rubber in weight. The main types of rubbers used are styrene-butadiene, acrylonitrile-butadiene and acrylic rubbers.

3.4. Insulation Corkboard

Insulation corkboard (ICB) is a product made of granules of cork expanded (steambacked) in autoclaves ($>$ 300 °C, $\sim$ 50 kPa) and agglomerated by pressure and temperature without any exogenous glue. Heat expansion results in an increase in cell volume of about 100 % and so the granules are compressed against each other and in the junctions the cells collapse. The blocks formed are then cut into planks which are then finished as required [57,58]. The main characteristics of ICB are listed in Table 3 [57,59].

Table 3. ICB average characteristics

Characteristic	Value
Density	110 kg/m^3
Thermal conductivity coefficient	0.045 W m^{-1} K^{-1}
Specific heat	1.9 MJ kg^{-1} K^{-1}
Thermal expansion coefficient	40×10^{-6}
Limit of pressure∗	10 000 kN/m^2
Permeability	3×10^{-5} m^{-1} h^{-1} Pa^{-1}
Tensile strength	0.05 MPa
Compressive strength at 10 %	0.25 MPa
Bending tension∗∗	1.6 kN/cm^2

∗ limit of pressure under which the material has an elastic behavior
∗∗ force for bending a material with a cross surface

Although such an agglomerate is produced with the lowest quality types of cork it has very favorable properties for some applications. Insulation corkboard is used for three applications: thermal insulation, acoustical absorption, an vibration insulation. The first application is the most important one. Acoustical corkboard,

for acoustic insulation is fabricated from specially selected raw materials, using granules of smaller size, lower density, and careful fine finish [59]. For vibration insulation higher densities are used.

Agglomerate for insulation cork board is exclusively made from cork granules (mainly from winter virgin cork) of a larger size (usually 5 – 20 mm) than those used in composition cork.

3.5. Other Specific Products

Most other cork products are made by similar processes as used for floor and wall covering cork products. For example, *filler and expansion joints* are specially designed to neutralize the expansion and contraction phenomena that can jeopardize concrete constructions, being an excellent protection against cracks, usually due to temperature variations. These materials can withstand continuous strain, whatever the moisture conditions are.

4. Uses

4.1. Stoppers

Good wines are made to grow and mature and they have to be protected from the environment but have to establish with it a slow and subtle relationship of exchange that only the cellular structure of cork can ensure. The friction coefficient of cork is high providing the ability to adhere glass due to the elastic force and the microscopic suction pads (cut cells).

Agglomerated corks are most commonly used in sparkling wines, but also in a sizable percentage of table wine (fast consumption wine). These corks are of very uniform size and density but are unlikely to provide the aging capabilities of the traditional cork stopper [1]. Colmated corks (pores covered with a mixture of cork powder are glue) are attractive low cost stoppers but are used only in common table wines.

The type of surface treatment applied on the cork stopper is determined by the type of wine and corkage cadence. The recommended moisture content for the stoppers in corkage is 6 – 8 % and the corker machine must press only 1 – 1.5 mm below the entry bore, avoiding compressions of more of 1/3 of the stopper diameter. When filling, an empty space level should be respected. After corking the bottle must be maintained in a stand-up position for some minutes to allow the dimensional recovery of the cork stopper.

In champagne corks, the base, where sealing is most critical, is usually made of two disks cut with the lenticels crossing the disks' thickness. Leakage problem is avoided because the lenticels of two or more disks do not connect [6].

4.2. Civil Construction

In buildings and other civil construction works, cork products may be used for thermal insulation and insulation, acoustic correction, floor covering, wall covering, false ceilings, and expansion joints:

- Thermal insulation – ICB
- Vibration insulation – ICB, corkrubber
- Acoustic correction – ICB, composition cork
- Floor covering – composition cork, cork rubber
- Wall covering – soft composition cork
- Expansion joints – composition cork

Some of these cork products may be used as composites with other construction materials, for example, ICB with polyurethane, composition cork with MDF and wood veneer, or may be covered by waxes, varnishes, and paints.

4.3. Industrial Applications

Some specific applications of cork in the industry include: cork slabs and sheets, pipe insulation, battery mold coating, cold storage insulation, machinery anti-vibration sheets, and storage tank insulation.

As expanded cork agglomerate (ICB) maintains its properties at very low temperatures, it has a wide range of applications in the refrigeration industry. The electrical industry uses corkrubber to stop leaks of transformer fluids. Similar products are also used in the manufacture of electrical switches, lightning apparatus and the like [2]. The ceramic industry uses cork

disks to smooth edges and to eliminate defects [2].

High density expanded agglomerate planks are one of the best means of insulating machines and damping vibrations [2, 59].

4.4. Automotive Industry

Cork makes good gaskets because it accommodates large elastic distortion and volume change and its closed cells are impervious to water and oils. Almost all the gaskets used in automobiles are made of corkrubber since this combination is easy to cut and due to its flexibility and resistance it can be used in gaskets with very narrow borders [2, 7].

4.5. Other Specific Uses

Other specific uses include: baseballs, cricket, hockey, and golf balls, golf clubs, plate mats, memoboards, musical instruments, polishing blocks, radioactive isotope capsules, ashtrays, bath mats, beehives, candle holders, glass coasters, cork paper, insulation in solid fuel rocket motor cases, desk pads, cork textiles, fishing rods rings and handles, fishing net floats, handicraft items, hot pads, jar lids, buoys, safety helmets, shoes, shuttlecocks, table-tennis bats, whistles, bungs for gun cartridges, calendars, etc.

5. Quality and Standardization

5.1. Standardization and Testing

Cork standardization was initiated in Portugal in 1957 with the establishment of the Technical Committee (TC-16) aiming at the development of standards for cork and industrial cork products, from raw material and terminology to finished products. Currently about 40 standards exist. ISO/TC-87 working since 1958 has the same aim and is responsible for the production of international standards. Portugal has the Chairmanship and the Secretariat of this TC. Within CEN (Standardization European Committee) there are three TC whose scope is directly related to agglomerated cork applications: CEN/TC-88 (thermal insulation), CEN/TC-99 (wall coverings) and CEN/TC-134 (resilient floor coverings). Several European Standards are now published concerning test methods and product specifications, supporting harmonized standards to be converted at a national level [60].

For a list of European and international standards see [60].

The requirements and *methods of testing* for the different cork products, as well as these products classification are described in both national and international and/or European standards [60].

5.2. Quality in Cork Production

There are a large number of productive units, certified according to the EN ISO 9000 series standards. Furthermore, the recent adoption and implementation of the International Code for Cork Stopper Manufacturing Practice reinforces and adds credibility to the quality guarantee process. There are also accredited laboratories, which specifically perform specific tests for cork products.

The Cork Quality Council has instituted a common grading standard in order to eliminate differences in different companies to categorize cork stoppers (visual characteristics). This standard is designed to be used by buyers. The criteria allow classification of cork stoppers in four grades (A, B, C, and D). The visual grades are designed to describe a range of corks of similar appearance.

6. Economic Aspects

6.1. Forestall Production

The cork oak adapts to soils practically unable of supporting any other economically viable crops, and prevents the degradation of these soils. The value of this tree should not be measured merely by the cork produced but rather by the whole aspects of agriculture, forestry, grazing, hunting, etc.

Because of particular climatic conditions (temperature, light, and rainfall) the cork oak established itself in the Western Mediterranean

basin. It is found in Northern Africa (Morocco, Algeria, Tunisia), in Southern France (e.g., Corsica), Italy (Sardegna), Spain and, above all in Portugal. Worldwide cork oak forests cover an area of about 2.2×10^6 ha [1,61]. A breakdown of cork forest surface and cork production per year worldwide is given in Table 4.

Table 4. World cork wood surface and production per year

Country	Surface, ha	Production, t
Portugal	660 000 – 730 000 (28 – 35 %)	170 000 – 185 000 (54 – 63 %)
Algeria	410 000 – 480 000 (12 – 21 %)	15 000 – 27 000 (4 – 12 %)
Spain	410 000 – 500 000 (18 – 23 %)	80 000 – 89 000 (22 – 29 %)
Morocco	340 000 – 400 000 (9 – 18 %)	13 000 – 18 000 (4 – 7 %)
France	100 000 – 110 000 (1 – 5 %)	5000 – 32 000 (1 – 11 %)
Italy	90 000 – 100 000 (4 – 10 %)	20 000 – 26 000 (5 – 6 %)
Tunisia	90 000 – 100 000 (3 – 6 %)	8000 – 9000 (1 – 2 %)

The world average cork productivity is around 150 kg ha^{-1} a^{-1}. An average tree (in age and size) may provide up to 100 kg per strip, with 7 – 11 kg/m^2 of debarked area. However, in the best cases these values can be much higher [1]. Although prices can vary considerably, as a reference, the price of good quality cork for stoppers in Portugal, in the campaign of 2000, was around 3 – 3.5 €/kg.

The undergrowth in corkwoods is rich in a wide variety of aromatic and medicinal plants and is also home to game of significant economic value. Acorns are used for swine food, wood is used for small pieces but mainly for production of firewood and charcoal [61].

6.2. Personnel and Companies

At present more than 900 companies are operating in the cork business in Portugal , mainly in the north of the country (Santa Maria da Feira). These firm employ about 15 000 persons, 85 % of the factories have less than 20 workers. In Spain about 280 companies are active in this field, mainly in Cataluña (province of Girona), Extremadura and Andaluzia, 120 of these companies produce cork stoppers.

In Portugal about 6000 workers work seasonably in cork extraction and this figure allows an extrapolation for other countries. It is estimated that there are about 150 000 people around the world, which are dependent on the cork sector, including production, processing, and marketing.

6.3. Consumption of Products and Markets

Taking the Portuguese industry as reference, the breakdown of cork consumption according to product is as follows [61]:

Natural cork stoppers	57 %
Agglomerated cork stoppers	11 %
Composition cork board	17 %
Insulation cork board	6 %
Other	8 %
Non finished products	1 %

As a reference (for the year 2000) a first quality cork stopper can cost as much as € 0.35 but a very wide range of prices exists for all types of cork stoppers, starting at less than € 0.05.

One square meter of simple cork floor covering can have a price as low as € 7 – 8, but also here a wide range is found.

Annual world consumption of cork products is given below [2]:

Cork stoppers for wine	13×10^9
Stoppers for champagne and sparkling wine	1.5×10^9
Insulation cork board	150 000 m^3
Floor tiles	10×10^6 m^2

The importance of the main sectors of cork products utilization, taking the Portuguese production as a reference, is as follows [4]:

Wine sector	61 %
Automobile sector	11 %
Civil construction	15 %
Other	13 %

6.4. Producers

Currently the cork industry is spread over different areas of activity:

– The preparer branch: selection and preparation of cork for the manufacturing industry

– The manufacturer branch: production of items by simply cutting the cork planks
– The crusher branch: production of cork granulates and raw materials for the cork board industry
– The blender branch: production of different types of cork boards

Cork products are manufactured both in small family companies and very large companies with thousands of employees. Some of the companies produce only one type of cork products, for example stoppers, and others manufacturing the entire variety of cork products.

Portugal and Spain process around 75 % of the world's cork production; France, Italy, and the North African countries produce together around 10 %, the remaining 15 % is scattered over different countries in Eastern Europe, Japan, Switzerland, Germany, UK, Latin America, India, and USA [1].

The most important producer is the Portuguese Amorim Group, which started its business in 1870. This group has 34 factories in Portugal and other countries with a total of more than 4000 workers and total sales in 1999 of about € 512 million, accounting for about 1/5 of the world cork processing capacity. This group produces cork components for almost all kinds of applications.

The second largest cork producing country is Spain, with the company Corchos de Mérida that produces 200×10^6 of cork stoppers (natural and granulates) with a transaction volume of € 18 million.

7. Environmental and Toxicological Aspects

7.1. Industrial Wastes

Most of the waste materials produced in cork processing are used in other types of cork products and therefore cannot be considered as an industrial waste. However, the cork processing still gives rise to real specific waste materials such as cork dust, cork boiling wastewaters, stoppers washing waste solutions, and steambaking condensates.

Cork powder waste (cork dust) is formed in different industrial cork processing operations and from different cork types. Worldwide cork powder production is around 50 000 t/a [62]. For studies on cork dust composition see [63].

By definition, cork dust is the group of cork particles with a size smaller than 0.25 mm [62]. Cork dust and cork fungi (from cork processing and manipulation) may be present in the air, even if extraction systems are installed in the factories. Prolonged occupational contact with these agents, if care with protection systems such as masks is not taken, may be related with an illness called suberosis.

Considerable quantities of cork dust are produced in grinding, granules cleaning and separation, cutting, sanding, and other finishing operations. More than 25 % of processed cork becomes cork dust [62]. Pure cork dust may be used in cork stoppers finishing operations and other specific applications [68, 69]. The rest is incinerated to produce steam (e.g., for ICB backing) or energy (drying, electricity etc.). This waste has a high heating value (15 – 28 MJ/kg). For ICB production 180 kg of cork dust are consumed per cubic meter [1] or 1.45 kg per kilogram ICB [70].

Steambaking Condensates. The ICB steambaking condensates are dark brown substances, which consist mainly of tannins and waxes. The amount of steambaking condensates formed annually is difficult to estimate but according to some calculations is about 1000 t/a.

Wastewater. The wastewater from cork processing amounts to 0.45 m^3 per ton boiled cork [1] (120 – 130 t/a world wide, mostly tannins).

7.2. Recycling and Treatments

Cork materials formed as waste in one operation can often be used in another operation of cork transformation. For example, the waste of cork from the cork stopper industry is used to produce cork granulates and cork powder from different origins is used to fill pores of some stoppers or is used as an energy source in cork factories [62] (see Section 7.1).

Cork grinding waste can be used as a substrate for plants with very good results [71] in germination and plant growing.

Used cork stoppers may be used as a raw material for the production of cork granulates or for the production of utilitarian wares (e.g., cutlery holders, table wares, pullers) according to a patented process [67].

7.3. Emissions

In the production and processing of some cork based products sometimes adhesives, varnishes, paints, glues, coatings, and other additives are used. Substances are formed or liberated and are emitted from the plants and eventually from finished products.

When urea – formaldehyde or phenol – formaldehyde resins are used as adhesives for cork materials, the finished products can potentially emit formaldehyde. The emission depends on the type of binding agent, catalyst, operational conditions, stabilization periods etc. (see → Formaldehyde; → Phenolic Resins; → Amino Resins). Studies carried out on cork materials showed that the level of formaldehyde emitted from cork board ranged from about 0.02 to 0.2 mg/100 g, formaldehyde emitted by natural cork ranged from 0.03 – 0.04 mg/100 g. The stabilization period (after manufacturing and before marketing) has almost no effect on this [72].

8. Legal Aspects

Protection of acorn production date back as far as the Costumes de Castelo Rodrigo e Castelo Melhor in 1209 in Portugal [2]. Further legislation was enacted during the following centuries to protect the cork tree, setting regulations for thinning, pruning, and stripping the cork. Current legislation covers the fields of forestland, cork oak, cork, and the cork oak farmer. The cork oak is protected from the moment of germination. A special license is required to fell trees. The same law requires the tree owners to take preventive measures, regarding to diseases, fire and damages by animals. Thinning is only permitted if the normal density is disrupted and the owner must notify the authorities. Pruning is only permitted from December to March. The tree may only be stripped for the first time when the trunk has a perimeter of 70 cm and extraction of reproduction cork is expressly prohibited if this cork is less than 9 years old [1, 2].

9. References

1. L. Gil, "Cortiça – Produção, Tecnologia e Aplicação", Ed INETI, Lisbon, 1998.
2. M. A. Oliveira, L. Oliveira: *The Cork,* Ed. Corticeira Amorim, Rio de Mouro 2000.
3. L. J. Gibson, M. F. Ashby: *Cellular Solids. Structure and Properties,* Pergamon Press, Oxford 1988.
4. W. Liese, H. Gunzerodt, N. Parameswaran: *Biological Alterations of Cork Quality Affecting its Utilization,* Ed. Instituto Produtos Florestais, 1983.
5. L. Gil, "História da Cortiça", Ed. APCOR, Santa Maria de Lamas, 2000.
6. "The art of cork", Ed. ICEP/AIECN/AIEC, Lisboa.
7. M. F. Ashby: *Cork in Concise Encyclopedia of Wood and Wood Based Materials,* Pergamon Press, 1989, pp. 66 – 69.
8. H. Pereira, *Wood Sci. Technol.* **26** (1992) 259 – 269.
9. H. Pereira, *Wood Sci. Technol.* **22** (1988) 211 – 218.
10. E. Conde, M. C. García-Vallejo, E. Cadahía, *Wood Sci. Technol.* **33** (1999) 271 – 283.
11. P. Sitte, *Protoplasma* **54** (1962) 555 – 559.
12. A. M. Gil, M. Lopes, J. Rocha, C. Pascoal Neto, *Int. J. Biol. Macromol.* **20** (1997) 293 – 305.
13. P. E. Kolattukudy, *Recent Adv. Phytochem.* **11** (1977) 185 – 246.
14. M. F. P. V. S. Bento, "Estudo dos Componentes da Cortiça e da Estrutura da Suberina por Técnicas Espectrométricas", Universidade Técnica de Lisboa, ISA, Ph. D. Thesis, 1997.
15. A. J. V. Marques, "Isolamento e Caracterização Estrutural da Lenhina da Cortiça de *Quercus Suber* L.", Ph. D. Thesis, Universidade Técnica de Lisboa, ISA, 1998.
16. N. Parameswaran, W. Liese, H. Günzerodt, *Holzforschung* **35** (1981) 195 – 199.
17. J. Carvalho: *A Text book of course of F.S.E., Técnicos para a Indústria Cortiçeira,* Universidade Técnica de Lisboa, IST Lisboa, 1987.

18. J. Graça, H. Pereira, *Holzforschung* **51** (1997) 225 – 234.
19. E. Seoane, E. Villacorta, M. J. De La Villa, *Anal. Quím.* **67** (1971) no. 3, 329 – 334.
20. B. Rodríguez-Miguéne, I. Ribas Marques, *Anal. Quím.* **68** (1972) no. 11, 1301 – 1306.
21. P. J. Holloway, *Chem. Phys. Lipids* **9** (1972) 158 – 170.
22. C. Agullô, E. Seoane, *Chem. Ind. (London)* **17** (1981) 608 – 609.
23. M. Arno, M. C. Searra, E. Seoane, *Anal. Quím. C* **77** (1981) 82 – 86.
24. P. J. Holloway, *Phytochemistry* **22** (1983) 495 – 502.
25. N. Cordeiro, N. M. Belgacem, A. Gandini, C. Pascoal Neto, *Bioresource Technol.* **63** (1998) 153 – 158.
26. N. Cordeiro, M. N. Belgacem, A. J. D. Silvestre, C. Pascoal Neto, A. Gandini, *Int. J. Biol. Macromol.* **22** (1998) 71 – 80.
27. L. Gil, *Silva Lusitana* **3** (1995) no. 1, 73 – 84.
28. A. Guillemonat, *Ann. Fac. Sci. Marseille* **30** (1960) 43 – 54.
29. A. Guillemonat, J. C. Trainard, *Bull. Soc. Chim. France* (1963) 142 – 144.
30. E. Agullô, E. Seoane, *Anal. Quim.* (1982) 78e, 389 – 393.
31. J. Graça, H. Pereira, *J. Wood Chem. Technol.* **18** (1998) no. 2, 207 – 217.
32. A. Björkman, Svensk, *Papperstid.* **59** (1956) 477 – 485.
33. W. Zimmermann, H. H. Nimz, E. Seemüller, *Holzforschung* **39** (1985) 45 – 49.
34. C. Pascoal Neto, N. Cordeiro, A. Seca, F. Domingues, A. Gandini, D. Robert, *Holzforschung* **50** (1996) 563 – 568.
35. A. J. V. Marques, H. Pereira, D. Maier, O. Faix, *Holzforschung* **48** (1994) 43 – 50.
36. A. J. V. Marques, H. Pereira, D. Maier, O. Faix, *Holzforschung* **50** (1996) 393 – 400.
37. A. J. V. Marques, H. Pereira, D. Maier, O. Faix, *Holzforschung* **53** (1999) 167 – 174.
38. C. Pascoal Neto, D. Evtuguin, D. Robert, *J. Wood Chem. Technol.* **14** (1994) no. 3, 383 – 409.
39. M. Lopes, C. Pascoal Neto, D. Evtuguin, A. J. D. Silvestre, A. Gil, N. Cordeiro, A. Gandini, *Holzforschung* **52** (1998) 146 – 148.
40. A. Asensio, *Carbohydr. Res.* **161** (1987) 167 – 170.
41. A. Asensio, *Carbohydr. Res.* **165** (1987) 134 – 138.
42. A. Asensio, *Can. J. Chem.* **66** (1988) 449 – 450.
43. A. Asensio, *J. Nat. Prod.* **51** (1988) no. 3, 488 – 491.
44. H. Pereira, *Boletim Instituto Produtos Florestais, Cortiça* **483** (1979) 259 – 264.
45. J. L. B. López, G. G. Curbera, I. Ribas Marques, *Anal. Fis. Quím.* **62-B** (1966) 865 – 869.
46. E. Seoane, E. Villacorta, *Anal. Quím.* **66** (1970) 1015 – 1016.
47. E. Seoane, E. Villacorta, M. J. De La Villa, *Anal. Quím.* **67** (1971) 329 – 334.
48. J. Fernandez-Salgado, I. Ribas-Marques, *Acta Científica Compostelana* **IX** (1972) no. 3 – 4, 139 – 144.
49. J. Fernandez-Salgado, I. Ribas-Marques, *Anal. Quím.* **70** (1974) no. 4, 363 – 366.
50. S. K. Talapatra, D. K. Pradhan, B. Talapatra, *Ind. J. Chem.* **16 B** (1978) 361 – 365.
51. A. Patra, S. K. Chaudhuri, S. K. Panda, *J. Nat. Prod.* **51** (1988) 217 – 220.
52. M. L. Mata, "Identificação dos Constituintes Químicos do Pó da Cortiça e sua Valorização por Hemissíntese", M.Sc. Thesis, FCT-UNL, Lisboa, 1991.
53. F. Mata, V. Marques, H. Pereira, *Boletim Instituto Produtos Florestais, Cortiça* **569** (1986) 68.
54. M. E. Rosa, H. Pereira, M. A. Fortes, *Wood Fiber Sci.* **22** (1990) no. 2, 149 – 164.
55. F. Cumbre, F. Lopes, H. Pereira, *Wood Fiber Sci.* **32** (2000) no. 1, 125 – 133.
56. K. C. Fugelsang, D. Callaway, T. Toland, C. J. Moller, *Practical Winery & Vineyard* Jan. – Feb. (1997) 50 – 55.
57. H. Pereira, E. Ferreira, *Mater. Sci. Eng. A* **111** (1989) 217 – 225.
58. L. Gil, *Wood Sci. Technol.* **30** (1996) 217 – 223.
59. H. Medeiros, ABC Insulation Corkboard, Ed. JNC, Porto.
60. M. F. Bicho, L. Gil, "*Cortiça – Guia Normativo*", Ed. IPQ/CTCOR. Lisbon 1999.
61. "The Cork oak and Cork", Ed. DGDR, Lisboa, 2000.
62. L. Gil, *Biomass and Bioenergy* **13** (1997) no. 1/2, 59 – 61.
63. L. Gil, J. Santos, M. I. Florêncio, *Boletim Instituto Produtos Florenstais-Cortila* **575** (1986) 255 – 261.
64. E. Moiteino, "Aplilalões de triterperiódes dos Condensados Negios da Cortila em Quimila Firne", M. Se Thesis, FET-UNL, 1992.
65. EP 651 760, 1994 (E. Moiteino, M. R. Tavanes, M. Y. Mareloeunto); *Chem. Abstr.* **122** (1995) 187820.

66. E. Moiteino, F. Justino, R. Tavanes, M. Y. Marelo Curto, M. I. Florêncio, M. S. Y. Naseimento, M. Pedro, F. Cequeira, M. M. Pinto, *J. Nat. Prod.* **64** (2001) 1273 – 1277.
67. Pt 102 013, 1997 (L. Gil).
68. L. Gil, *Ingenium* **54** (1991) 58 – 71.
69. L. Gil, *Ingenium* **48** (1991) 54 – 61.
70. L. Gil, *The European Congress on Renewable Energy Implementation*, Athens (1997) 705 – 706.
71. A. Pes, *Boletim Instituto Produtos Florestais-Cortiça* **471** (1978) 13 – 17.
72. L. Gil, N. Maurício, G. Cáceres, *Holz als und Werkstoff* **58** (2000) 47 – 51.

Combinatorial Methods in Catalysis and Materials Science

Oliver Brümmer, Symyx Technologies, Inc., Santa Clara, CA 95051, USA

Keith A. Hall, Symyx Technologies, Inc., Santa Clara, CA 95051, USA

Margarete Leclerc, Symyx Technologies, Inc., Santa Clara, CA 95051, USA

Silvia E. Lee, Symyx Technologies, Inc., Santa Clara, CA 95051, USA

Vince Murphy, Symyx Technologies, Inc., Santa Clara, CA 95051, USA

1. Introduction 2
2. Enzyme Mimetics and Metal Complexes as Enzyme Mimetics 4
2.1. Combinatorial Libraries of Catalysts as Enzyme Mimetics 4
2.2. Combinatorial Synthesis Enzyme Mimetics 5
2.2.1. Hydrolytically Active Metal Complexes 5
2.2.2. Evolutionary Solid-Phase Synthesis of Oxygenase Mimics 5
2.2.3. Libraries of Organic Acylation Catalysts 6
3. Combinatorial Catalysis in Asymmetric Synthesis 6
3.1. Combinatorial Catalyst Libraries in Enantioselective Additions of Dialkylzinc Reagents 6
3.2. Ligands for the Lewis Acid-Catalyzed Asymmetric Aza Diels–Alder Reaction 7
3.3. Divergent Ligand Synthesis for Enantioselective Alkylations 8
3.4. Chiral Phosphine Ligands for Asymmetric Hydrogenation 8
3.5. Asymmetric Reactions Catalyzed by Schiff Base-Type Ligands: the Positional Scanning Approach . . . 9
3.6. Identification of Novel Catalysts for Asymmetric Epoxidations by the Positional Scanning Approach 9
4. Multidimensional Combinatorial Screening for the Identification of Novel Catalysts 10
4.1. Catalyst Discovery and Optimization Using Catalyst Arrays 10
4.2. Parallel Array Screening for Catalyst Optimization Using Discovery and Focused Ligand Libraries . . . 11
5. One-Pot, Multi-Substrate Screening 12
6. Combinatorial Approaches to Olefin Polymerization Catalysts . . 12
7. Combinatorial Inorganic Catalysis 15
7.1. Combinatorial Libraries of Homogeneous Polyoxometalate-Based Catalysts 15
7.2. Combinatorial Libraries and High-Throughput Screening of Heterogenous Polyoxometalate Catalysts 16
8. Combinatorial Heterogeneous Catalysis 16
8.1. Oxidative Dehydrogenation of Ethane 17
8.2. Oxidative Dehydrogenation of Propane 19
8.3. Catalytic Oxidation of CO and Reduction of NO 20
9. Combinatorial Electrocatalysis . . 20
9.1. Electrocatalysts for Fuel Cells . . . 20
9.2. Combinatorial Electrosynthesis . . 21
10. High-throughput Screening Tools for Catalysis 22
11. Combinatorial Solid-State Materials Science 23
11.1. Materials Library Synthesis 23
11.1.1. Vapor Deposition Techniques 23
11.1.2. Alternative Library Synthesis Techniques 25
11.2. High-Throughput Screening for Materials 26
11.2.1. Optical Screening 27
11.2.2. X-Ray Characterization 27
11.3. Applications of Materials 27
11.3.1. Superconductivity 27
11.3.2. Ferromagnetic Semiconductors . . . 28
11.3.3. Magnetoresistant Materials 28
11.3.4. Dielectric and Ferroelectric Materials 29
11.3.5. Luminescent Materials 30
12. Organic Materials and Polymers . 31
12.1. Schiff Bases for Nonlinear Optical (NLO) Materials 31
12.2. Artificial Receptors for Small Organic Molecules 32

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.f07_f02

12.3. New Materials for the Separation of Enantiomers 32
12.4. Molecular Imprinting 33
12.5. Polymer Topologies and Functionalization 34
13. Summary and Outlook 35
14. References 36

1. Introduction

The conventional or "one-at-a-time" approach for discovery and process or performance optimization in the areas of biopharmaceuticals, materials science, and catalysis has traditionally been a time-consuming and expensive endeavor and often relies upon a large number of iterative steps. This is because technological problems are largely multivariable in nature and it is seldom possible to predict a priori which of the variables are most important in arriving at a solution. Increasing competition in the chemical and pharmaceutical industries has necessitated faster and more efficient means of identifying new materials and formulations. This drive towards increased research productivity was first encountered in the pharmaceutical industry, where long development times and high research costs resulted in the introduction of new R & D approaches designed to accelerate the drug-discovery process. One such approach is the use of high-throughput methods for rapidly synthesizing and screening large numbers of potential drug candidates for their biological activity. Most major drug manufacturers now utilize this so-called combinatorial chemistry (→ Combinatorial Chemistry) as an integral part of their research and development programs [1–8].

In the broadest terms, the combinatorial approach involves the design and high-throughput synthesis and screening of arrays containing large numbers of chemical or biochemical materials or formulations. Arrays (often referred to as libraries) are designed to span a particular variable or phase space of interest and are screened for one or more physical properties or chemical parameters of interest. The term "high-throughput" can be somewhat subjective, and generally depends upon the particulars of the chemical problem under study and rates of conventional experimental throughput. For example, in pharmaceutical research high-throughput can refer to the synthesis and screening of hundreds of thousands of drug candidates per year. In other areas this number may be higher or considerably lower but still well beyond the capabilities of conventional research.

The combinatorial approach has found application not only to the discovery and optimization of new materials or formulations but also existing (and possibly commercialized) ones as well. The path towards discovery, optimization, and eventual commercialization of such materials or formulations is depicted in Figure 1. The combinatorial approach begins with high-throughput screening, which can be divided into various stages. During the first stage, referred to as primary screening, many experimental combinations are surveyed as quickly as possible for one or more properties or performance characteristics to eliminate the vast majority of combinations. The figures of merit chosen to analyze each combination may be measured directly or by inference from another related property. To accommodate the high-throughput of a primary screen it is often necessary to tolerate certain compromises in the quality or precision of the data obtained. Materials or formulations that meet certain performance criteria during the stage of primary screening (hits) are then investigated more thoroughly in a secondary screen. Typically, secondary screening involves preparation of libraries of materials or formulations on a larger scale under more controlled conditions, and an analysis that is more rigorous and discriminating. As a result, the number of experiments performed in the secondary screen is typically smaller than in primary screening, the trade-off being throughput versus precision of data. Candidates that perform well in secondary screening are then selected for further studies (tertiary screening), for example, to optimize performance or establish structure – activity relationships. In this stage of the screening process, arrays are constructed by using phase space parameterized around key experimental variables identified during secondary screening. The process is iterated until materials or formulations with the desired properties or performance criteria are discovered.

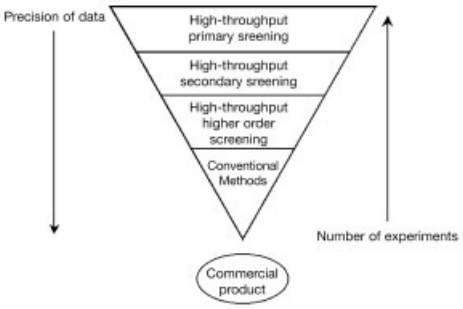

Figure 1. Process of discovery and optimization of a commercial material or process by combinatorial methodologies

For example, a primary screen for a new catalyst system might be based solely upon activity. The most active elements of the libraries screened would then be chosen for secondary screening, in which catalyst activity and another property, such as selectivity, are determined. Those catalysts that have the best performance based on activity and selectivity in the secondary screen might then be singled out for further study using focus libraries to optimize catalyst performance. The focus libraries might include structural variations of the catalysts that emerged from the secondary screen or a more expansive range of experimental conditions than were initially used.

The materials emerging from the combinatorial process of primary, secondary, and higher order screening are then studied further by conventional methodologies employing standard laboratory reactors and characterization tools towards the goal of producing a new or improved commercial product (Figure 1). Exactly what stage conventional, one-at-a-time methods enter the process of discovery and optimization and what these methods entail depends largely upon the material or process being studied. Conversely, there may be practical limitations to the scope and applicability of combinatorial methods. For example, the implementation of a new catalyst discovered by combinatorial approaches may require conventional, one-reaction-at-a-time scale-up studies in batch reactors and pilot plants prior to implementation in a commercial process. In this example, combinatorial approaches are a powerful research tool used in advance of or in conjunction with more conventional methods of discovery and optimization.

The efficient execution of the combinatorial approach requires the use of rapid serial or parallel synthesis and analytical methods augmented by the judicial use of automation, miniaturization, and software and database technologies. Miniaturization and automation are often critical to accommodate the experimental scale and relieve the tedium imposed in the creation of highly dense libraries. The use of hardware and software is also important, not only to enable automation, but also as an aid in library design and analysis. Moreover, the accumulation of a tremendous amount of data related to the synthesis and screening of libraries must be efficiently managed. To this end numerous database and data-mining technologies such as bio- and chemoinformatics have been spawned [9]. The use of such data-managing technologies allows key and often nonobvious structure – function relationships to be understood, thereby providing a wealth of information for use in future experimental designs. Indeed, the ability to conduct experiments in a high-throughput fashion not only allows the scientist the freedom to explore well-constructed ideas but also engenders serendipity.

The development of combinatorial technology for the biopharmaceutical industry was initially leveraged by the availability of low-cost, high-performance computers, robotics, molecular modeling, and database software, as well as tools created for the clinical diagnostics, optics, and semiconductor industries. Similarly, the first application of the combinatorial approach to nonbiological systems was developed by using technology that has its origins in the semiconductor industry. Seminal work by Schulz et al. in 1995 demonstrated the ability to prepare semiconductor materials in an array format and sparked tremendous interest in this field [10]. However, certain concepts related to the synthesis and characterization of arrays of materials can be traced back to work published by Hanack in the early 1970s [11]. Since then, combinatorial methodologies have been applied to the discovery of optical and electronic materials, magnets, polymers, and catalysts for commodity, specialty, and fine-chemical synthesis. This has been made possible through the application and adaptation of technologies first established in the pharmaceutical and materials science industries. For example, research in combinatorial

catalysis has taken advantage of numerous solid- and solution-phase synthetic methodologies that have their origin in the pharmaceutical industry [12–15].

An increasing number of reviews address new methods in combinatorial chemistry and high-throughput synthesis and screening for chemical process development [16] and catalysis [17–22], and a glossary of frequently used terms has been published [23].

2. Enzyme Mimetics and Metal Complexes as Enzyme Mimetics

Enzymes have found widespread use in industry in a multitude of chemical reactions because of their excellent selectivities and high activities under relatively mild reaction conditions [24]. An ongoing field in research is the rational design and synthesis of numerous classes of low and high molecular mass synthetic enzyme catalysts (synzymes) that mimic, improve, substitute, or exceed a natural enzymatic function or structure.

2.1. Combinatorial Libraries of Catalysts as Enzyme Mimetics

One approach to identify enzyme-like new organic catalysts is by searching through a multitude of "randomly" generated polymeric systems [25]. One of the first applications of this concept in combinatorial catalysis was the development of phosphatase-like catalysts [26, 27] and a combinatorially developed polymeric reducing agent [28]. Another example is the dehydration of β-hydroxyketones [29]. A library of 1344 polymeric variants with an amide functionalization ranging up to 20 % was obtained by aminolysis of poly(acrylic anhydride) and sets of mixtures of three to four amines taken from a library of 11 amines (Figure 2).

Assuming that these polymeric catalysts possess both acid and basic functional groups in proper spatial orientation for catalyzing the dehydration of β-hydroxy ketones, the researchers spectroscopically monitored the appearance of the β-aryl α,β-unsaturated carbonyl chromophore of the product. About 1 of the polymeric catalysts revealed a significant rate enhancement ($k_{cat}/k_{uncat} = 920$).

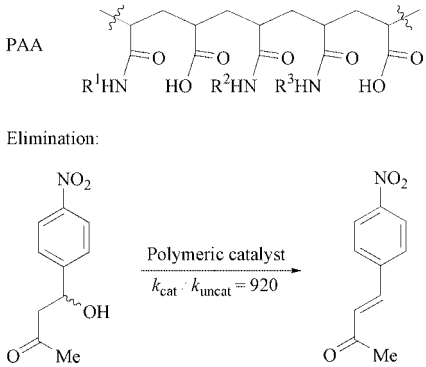

Figure 2. Example of a combinatorially functionalized polymeric catalyst based on poly(acrylic anhydride) facilitating the dehydration of β-hydroxyketones

2.2. Combinatorial Synthesis Enzyme Mimetics

2.2.1. Hydrolytically Active Metal Complexes

Enzyme mimics for activated ester or phosphate hydrolysis remain in the center of interest. Janda et al. synthesized and screened transition metal complexes as enzyme mimics that catalyze the hydrolysis of carboxylic esters [30]. The researchers extended their methodology towards the synthesis of the corresponding azacrown ether lanthanide complexes that were capable of hydrolyzing phosphate esters, a crucial chemical transformation in areas such as nucleic acid chemistry [31].

Using a split – pool approach (→ Combinatorial Chemistry), Berkessel and Hérault synthesized a 625-member library of undecapeptides on a solid support, which was assayed for catalytic activity in phosphate hydrolysis [32]. About 2500 beads were incubated with Cu^{II}, Zn^{II}, Fe^{III}, Co^{III}, Eu^{III}, Ce^{IV}, and Zr^{IV} solutions; only in the presence of Zr^{IV} catalytic hydrolysis activity was observed. [32]

2.2.2. Evolutionary Solid-Phase Synthesis of Oxygenase Mimics

The catalytic oxygen transfer activity of the enzyme Cytochrome P450 oxygenase inspired

Sames et al. to develop polymer-supported diamide complexes of manganese for catalytic oxygenation reactions [33]. Upon transformation into the corresponding manganese(III) complexes, the resulting catalysts were directly screened on-bead for the epoxidation of vinylbiphenyl (Figure 3). The first-generation library quickly showed that a number of structural variants led to respectable yields of the epoxide (74 % yield for the first-generation ligand). The second round identified a ligand which led to nearly complete consumption of vinylbiphenyl in 16 h (96 % yield). The third generation of ligands provided a cyclic system that exhibited the highest turnover rate of the entire assay; 5 mol % of this catalyst yielded 98 % of epoxide in 2 min, which translates into a rate enhancement by two orders of magnitude in comparison to the second-generation catalyst. The same catalysts were also used for the oxidation of ethylbenzene derivatives.

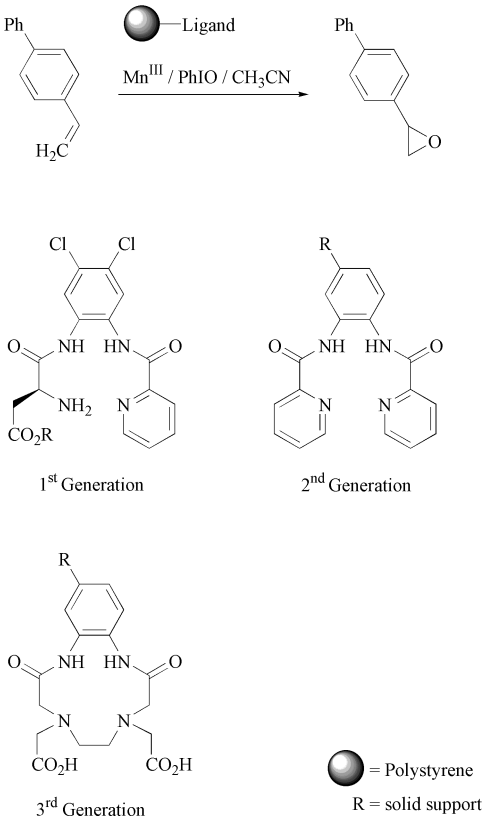

Figure 3. Evolution of diamide ligands for Mn^{III}-catalyzed oxygenation on a solid support

2.2.3. Libraries of Organic Acylation Catalysts

Mimics for acyl transferases require a concerted interplay between tethered functional groups, whereby the catalyst should possess a suitable base that is covalently tethered in close proximity and in the correct orientation relative to a nucleophilic center. Taylor, Morken et al. reported the preparation of a library of 3150 distinct nucleophilic acylation catalysts on approximately 7000 encoded polymer beads prepared by split – pool methods [34]. Temperature changes due to the exothermic catalytic acetylation of ethanol were monitored directly on the floating beads by IR thermography. The "hottest" 23 beads were then manually selected, and decoding revealed three novel structures as effective catalysts.

3. Combinatorial Catalysis in Asymmetric Synthesis

The importance and practicality of asymmetric synthesis as a tool for obtaining optically pure or enantiomerically enriched compounds have been fully acknowledged by chemists in the areas of synthetic organic, medicinal, agricultural, and natural products chemistry and their related industries [35]. Among asymmetric transformations, the most desirable and also the most challenging is catalytic asymmetric synthesis, in which one chiral catalyst molecule can create millions of targeted chiral product molecules in a process that has been termed "chirality multiplication". Thus, asymmetric catalysis has a significant economic advantage over stoichiometric asymmetric synthesis, and consequently the development of new methods for discovering asymmetric catalytic transformations is an important and emerging objective in chemistry. Among other approaches, divergent ligand synthesis strategies, in which an advanced ligand intermediate is transformed into a series of different chiral ligands, appear to be especially fruitful in this area. In most cases, the ligand bears the chiral information and therefore creates a discriminating chiral environment in close proximity to the active metal site of the catalyst. Asymmetric synthesis is an especially suitable area for

combinatorial catalyst discovery, since asymmetric catalysis is often mechanistically very complex, and critical variables in the parameter space to be evaluated for catalyst development and optimization are mutually dependent. Thus, by rapid parallel synthesis and metal complexation of chiral ligand libraries, coupled to automated or high-throughput screening, multidimensional problems in the discovery and optimization of catalysts may be efficiently addressed.

3.1. Combinatorial Catalyst Libraries in Enantioselective Additions of Dialkylzinc Reagents

The enantioselective addition of dialkylzinc reagents to aldehydes can be catalyzed by enantiopure aminoalcohols [36]. Ellman et al. reported the parallel solid-phase synthesis of a small library of chiral aminoalcohols, substituted 2-pyrrolidinemethanol ligands [37]. To anchor *trans*-4-hydroxy-L-proline methyl ester as a ligand scaffold to Merrifield resin, a parallel solid-phase synthesis approach involving a cleavable tetrahydropyranol linker was used. Subsequent nucleophilic addition of various Grignard reagents, followed by reductive *N*-alkylations, gave the desired aminoalcohols (Figure 4). On-bead screening of the asymmetric addition of diethylzinc to a variety of aromatic (up to 94 % *ee*) and aliphatic aldehydes (up to 85 % *ee*) yielded slightly inferior enantioselectivity compared to that of the corresponding free ligand in solution (Figure 4).

Gennari et al. used polymer-bound reagents [38], solid-phase extraction (SPE) [39], and polymeric scavengers [40] to speed up the synthesis of a library of 60 chiral Schiff base ligands that contain a set of different metal-binding sites (a phenol, an imine, and a secondary sulfonamine) [41]. Through a high-throughput screening approach, a ligand was identified that catalyzed the copper-mediated, enatioselective conjugate addition of diethylzinc to cyclic enones with up to 90 % *ee* (Figure 5).

In a related approach, Liskamp et al. developed a solid-phase synthesis of bis-sulfonamide ligands [42]. A library of these ligands was screened in a simultaneous substrate-screening

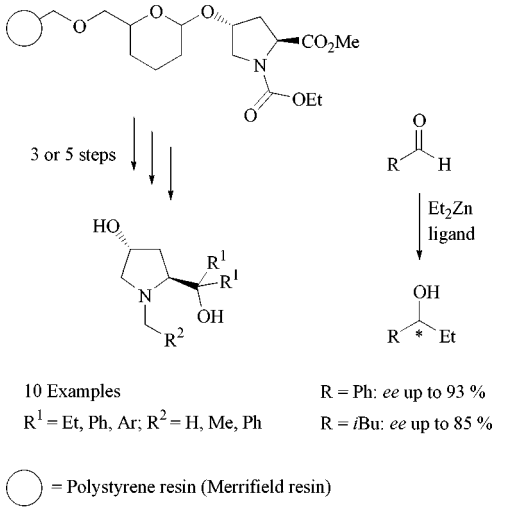

Figure 4. Parallel solid-phase synthesis of pyrrolidine-type catalysts and their screening in the enantioselective addition of diethylzinc to aldehydes

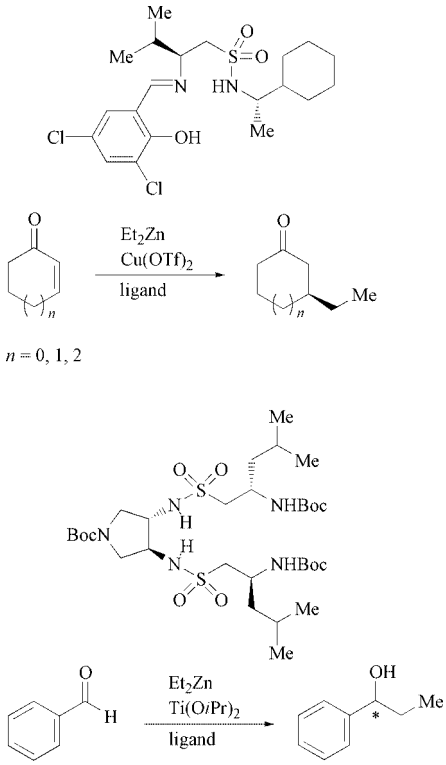

Figure 5. Ligands for enantioselective addition of diethylzinc, generated from combinatorial solid-phase synthesis

procedure for the ability to enantioselectively catalyze the $Ti(OiPr)_4$-mediated addition of di-

ethylzinc to aldehydes. The best solid-phase candidate was resynthesized in solution and gave enatioselectivities of up to 66 % (Figure 5).

3.2. Ligands for the Lewis Acid-Catalyzed Asymmetric Aza Diels – Alder Reaction

The asymmetric aza Diels – Alder reaction provides an effective route to optically active nitrogen-containing heterocycles [43]. Asymmetric catalysis of imine dienophiles has remained elusive, and successful asymmetric induction in the aza Diels – Alder reaction heavily relied on auxiliary-based methodology.

Kobayashi et al. reported a novel class of chiral zirconocene complexes that catalyze the aza Diels – Alder reaction of aldimines with Danishefsky's diene to afford the corresponding piperidine derivatives in high yields with high enantioselectivities [44]. In a solid-phase approach, polymer-supported (*R*)-1,1′-binaphthols (BINOLs) were synthesized and rapid catalyst optimization could be achieved (Figure 6). After synthesizing the most promising ligands in solution and further optimization of the inorganic precursor, a Zr complex was finally identified as the best catalyst, giving 68 % yield and 94 % *ee*.

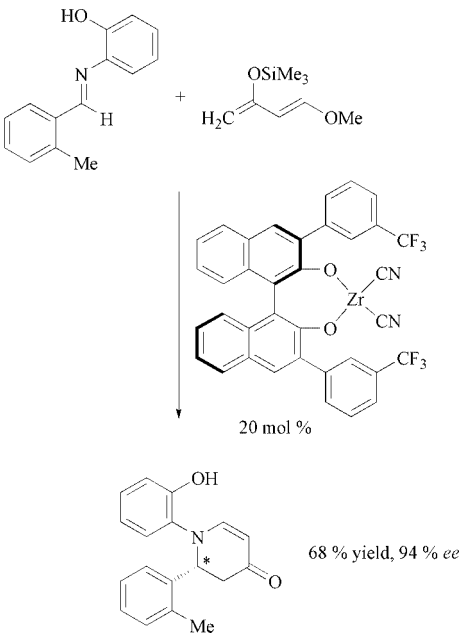

Figure 6. Solid-phase approach for the identification of a catalyst for the asymmetric aza Diels – Alder reaction

3.3. Divergent Ligand Synthesis for Enantioselective Alkylations

In 1998, Burgess et al. highlighted in several examples the value of divergent ligand synthesis strategies for the preparation of focused ligand libraries and the use of ligand screening [45]. By definition, in divergent syntheses of chiral ligands, an optically pure material is produced and then subsequently used to prepare a class of structurally similar ligands.

Based on L-serine, a chiral building block (chiral synthon = chiron) was prepared as a key intermediate that was then converted into a family of 13 novel chiral phosphine oxazoline ligands using several different routes and employing conventional organic solution synthesis [45]. It was shown that the divergent ligand synthesis approach was very suitable for efficiently accessing this ligand class with good diversity in terms of steric bulk and/or electronic pertubation (Figure 7).

The ability of these ligands to asymmetrically catalyze the addition of dimethyl malonate to 1,3-diphenylprop-2-enyl acetate was investigated [46]. The best of these ligands gave a palladium complex that catalyzes the allylic alkylation with up to 94 % *ee*.

In a different modular approach to generate chiral catalysts for the same transformation, Gilbertson et al. synthesized a phosphine oxazoline ligand library from amino acids and phosphino carboxylate building blocks [47, 48]. The best of these ligands gave a palladium complex that catalyzes the addition of dimethyl malonate

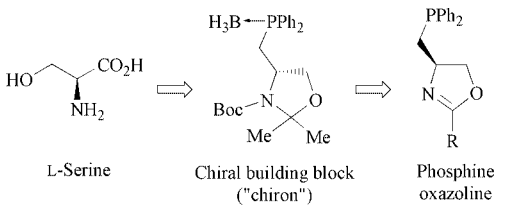

Figure 7. Divergent approach to phosphine oxazoline ligands
Boc = *tert*-butyloxycarbonyl, R = alkyl, aryl

to 1,3-diphenylprop-2-enyl acetate in 99 % yield and with 98 % enantiomeric excess.

Kobayashi et al. developed a solid-phase strategy to synthesize novel chiral pentamine ligands for the α-alkylation of lithium amide enolates, and obtained enantioselectivities of up to 84 % [49].

3.4. Chiral Phosphine Ligands for Asymmetric Hydrogenation

In homogeneous catalysis, phosphines are a particularly important ligands in late transition metal catalyzed processes such as hydrosilylation, hydroformylation, carbonylation, olefin dimerization, and hydrogenation [50]. For example, industrially important hydrogenations are involved in the synthesis of *l*-dopa (Monsanto) and Merck's novel HIV protease inhibitor Indinavir marketed currently as Crixivan [51].

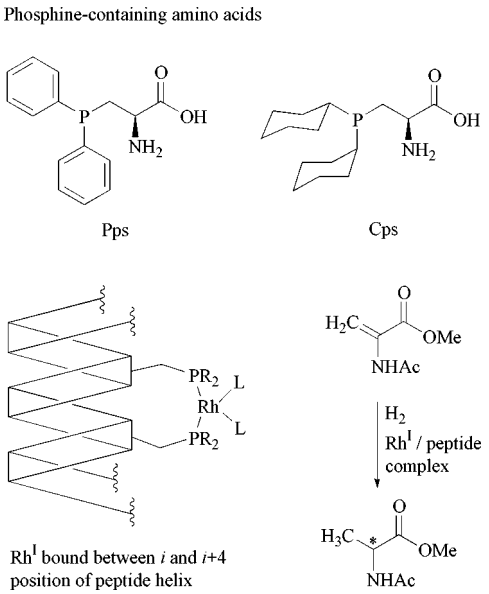

Figure 8. Screening of a library of support-bound α-helical peptide/phosphine Rh^I catalysts in enantioselective hydrogenation reactions of prochiral methyl-2-acetamidoacrylate to *N*-acetylalanine methyl ester

Gilbertson et al. synthesized a 63-member library of phosphine-containing amino acid ligands for the Rh^I-catalyzed hydrogenation of methyl 2-acetamidoacrylate to the corresponding amino acid derivative *N*-acetylalanine. Structurally unique chiral tetra- or pentapeptides were synthesized by combining the novel amino acids dicyclohexyl- and diphenylphosphine serine (Cps and Pps, respectively) with two or three hydrophobic amino acids, giving access to ligands that were used for the catalytic asymmetric hydrogenation (Figure 8, see above). Enantiomeric excesses of up to 38 % were obtained [52].

3.5. Asymmetric Reactions Catalyzed by Schiff Base-Type Ligands: the Positional Scanning Approach

A modular approach based on well-established functional building blocks can be used for the rapid discovery of novel catalysts [53,54]. The philosophy of this positional scanning approach for identifying a chiral catalyst consisting of three modular subunits, for example, is to optimize each of the three modular subuntits while keeping the other two constant. Common to the positional scanning approach (iterative optimization approach) of the modules comprising the ligand structure are the two assumptions (1) of an independent and additive influence of each of the building blocks, and (2) of the absence of negative cooperative effects during their systematic variation. In addition, the positional scanning approach, especially when coupled with automated screening, not only offers unique opportunities for the identification of substrate-specific catalysts (fine tuning) but also permits the discovery of ligands that possess unusual properties and which might otherwise elude detection.

In 1996, Snapper and Hoyveda reported the diastereo- and enantioselective addition of trimethylsilyl cyanide (TMSCN) to epoxides catalyzed by titanium Schiff base complexes to yield chiral β-cyanohydrins [53]. Chiral dipeptidic hemisalen ligands of the general structure SB-AA_2-AA_1-GlyOMe (SB = Schiff base or hemisalen, AA_i = amino acid) were synthesized in a modular fashion from various amino acid and aldehyde building blocks by using parallel solid-phase synthesis techniques. After cleavage from the solid support, complexation with a titanium alkoxide precursor, and parallel screening of the chiral Lewis acids for efficiency

and enantioselectivity, a potent catalytic system for ring opening of cyclohexene oxide with 89 % enantiomeric excess was discovered (Figure 9).

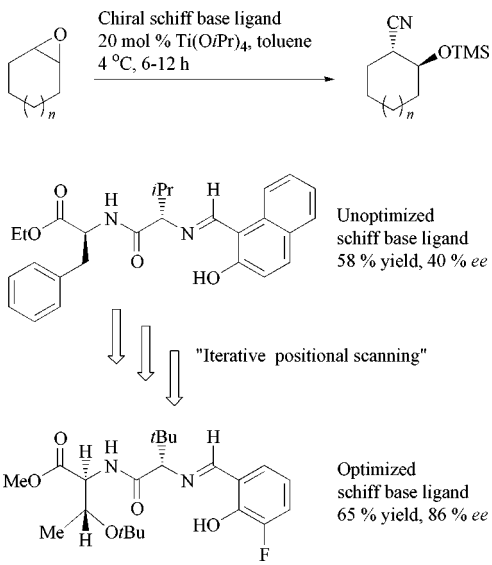

Figure 9. Discovery of a new Schiff base catalyst for the Ti-catalyzed enantioselective addition of trimethylsilyl cyanide (TMSCN) to *meso*-epoxides by an iterative optimization approach using positional scanning

Subsequent work by the same group demonstrated that screening did not require cleaving the Schiff base ligand from the solid support, which significantly accelerated the optimization process [55].

The positional scanning approach was used to identify novel catalytic systems for the asymmetric hydrocyanation of imines (Strecker reaction) [54], for enantioselective additions of cyanide to imines [56, 57], for Cu-catalyzed conjugate addition of alkylzinc reagents to cyclic enones [58], for Zr-catalyzed addition of diethylzinc to imines [59], and for Cu-catalyzed allylic substitutions [60].

3.6. Identification of Novel Catalysts for Asymmetric Epoxidations by the Positional Scanning Approach

In 1999 Jacobsen's group disclosed the discovery of novel chiral catalyst for the epoxidation of *trans*-β-methylstyrene (TBMS) [61] Out of 5760 possible metal – ligand combinations, a catalytic system comprising a chiral ligand and $FeCl_2$ was identified that produced moderate enantioselectivities (15 – 20 % *ee*) for the target reaction.

Yamamoto et al. describe the vanadium-catalyzed asymmetric epoxidation of allylic alcohols using α-amino acid-based hydroxamic acid ligands [62] To optimize the ligands, again the iterative positional optimization approach was used. The asymmetric epoxidations were performed in the presence of 1 mol % of $[VO(OiPr)_3]$ and 1.5 mol % of the best ligand, giving the corresponding chiral epoxy alcohols in 58 – 99 % yield and 76 – 96 % *ee*.

4. Multidimensional Combinatorial Screening for the Identification of Novel Catalysts

Conventional research in a "one-at-a-time" or serial fashion provides thorough quality control of the samples entering the screening process to insure that accurate data are obtained. Combinatorial methods that involve, for example, split – pool synthesis [23] are much faster and make the preparation of relatively large numbers of compounds feasible but often lack control over the purity of the compounds entering the assay screen. To combine the two extremes outlined above, methods have been developed that are based on parallel or array syntheses or screening in a spatially addressable format, usually with one compound per well, coupled to automated screens. This approach facilitates a relative high level of quality control and a significantly enhanced throughput relative to sequential methods. Thus, multidimensional problems in the discovery and in the optimization of catalysts can be efficiently addressed by coupling parallel ligand synthesis with automated or high-throughput multidimensional screens (e.g., metal precursors, ligands, solvents, and reactant ratios). An integrated instrumental set-up for parallel high-throughput catalyst screening should encourage researchers to investigate a broader variety of systems, some of which may be regarded as too unusual to warrant testing in a sequential fashion.

4.1. Catalyst Discovery and Optimization Using Catalyst Arrays

Intramolecular C–H insertion of metal carbenes generated by catalytic decomposition of diazocarbonyl compounds is a facile methodology for carbon – carbon bond formation. Burgess et al. used a highly parallel screening approach for the optimization of ligand – metal – solvent combinations for this reaction (Figure 10) [63]. Using the optimized set of reaction conditions ($AgSbF_6$, THF, 25 °C, 24 h), the desired product was obtained in 75 % yield and in a diastereomeric ratio of 3.5:1.

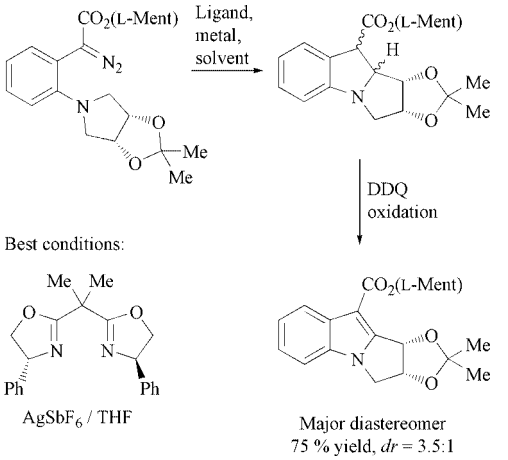

Figure 10. Discovery and optimization of a new catalyst for the intramolecular C–H insertion reaction of Rh carbenes from α-diazo esters for the synthesis of indolyl derivatives by high-throughput catalyst screening
L-Menth: L-menthyl; *dr* = diastereomeric ratio, DDQ = 2,3-dichloro-5,6-dicyano-1,4-benzoquinone

Using a small library of a novel phosphine oxazoline ligand class obtained by a divergent ligand synthesis approach [45, 64], parallel assaying was employed to evaluate and optimize the application of these ligands by Burgess et al. The ligands were screened simultaneously in a custom parallel reaction block with a 27- or 34-well format. The asymmetric alkylation of 1,3-diphenylpropenyl acetate with malonate was chosen as a classical test reaction (Figure 11) [46] Allylic alkylation with an enantiomeric excess as high as 94 % was observed.

The same concept was used to identify stereoselective catalysts for the diastereoselective reductive aldol reaction [65], for metal-catalyzed addition of amines to C–C double bonds [66], for Lewis acid catalyzed aza Diels – Alder reactions [67], for hydrosilylations [68], and for the metal-catalyzed reaction of enamines with Michael acceptors [69].

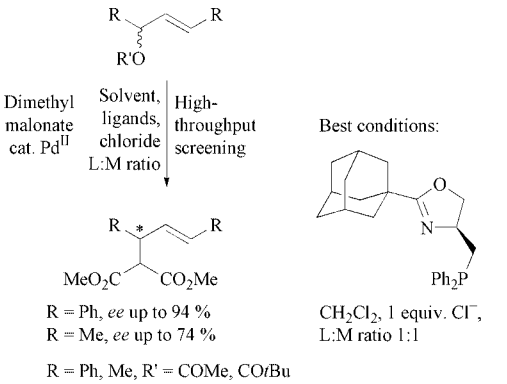

Figure 11. High-throughput screening of phosphine oxazoline ligands in Pd-catalyzed asymmetric allylic alkylation reactions

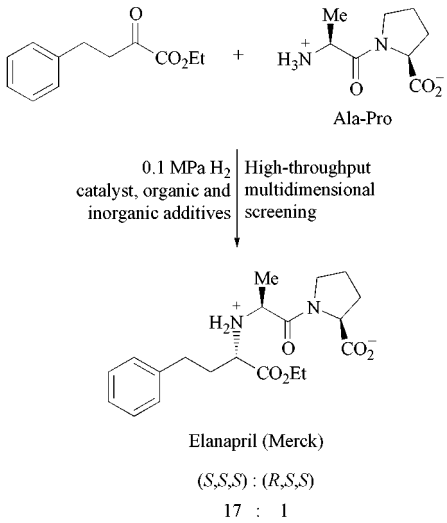

Figure 12. High-throughput screening of the Raney nickel catalyzed reductive amination between ketoester and the dipeptide Ala-Pro for Elanapril synthesis

Huffman and Reider from Merck Research Laboratories reported the discovery of improved conditions for the diastereoselective reductive amination in the large-scale synthesis of the ACE inhibitor Elanapril [70]. A multidimen-

sional screening approach was used to assay the effects of various heterogenous catalyst formulations and a variety of additives (Figure 12).

It was found that acetic acid and potassium fluoride as additives improved the (*S*,*S*,*S*) : (*R*,*S*,*S*) selectivity with Raney nickel in ethanol from 11:1 to 17:1. Simultaneously varying multiple reaction parameters reportedly led to increased diastereoselectivity and isolated yield improvement [90 %, (*S*,*S*,*S*) + (*R*,*S*,*S*) isomers].

4.2. Parallel Array Screening for Catalyst Optimization Using Discovery and Focused Ligand Libraries

In analogy to the pharmacological model of drug discovery, a large parent set of ligands can be screened for lead compounds. After screening of the ligand library and discovering successful hits, smaller, more focused daughter libraries can be created based on ligand structure – activity relationships, which result in a higher hit rate because of the closer diversity space. This screening approach is useful for the optimization of reaction parameters, especially when essential reaction parameters of the targeted reaction such as preferred ligand class, mechanism of the catalytic cycle, or metal precursor are already known.

Parallel solution-phase synthesis techniques and parallel-array assaying facilitated the identification and optimization of new chiral dirhodium(II) carboxylates for enantiselective Si–H insertion reaction of diazoesters with silanes [71]. A systematic search through the Available Chemical Directory (ACD) resulted in a set of 2000 useful alternative chiral carboxylic acid ligands which were divided into 80 different structural clusters [72]. Representative chiral carboxylic acid ligands from each of the 80 clusters were then used to prepare dirhodium(II) carboxylate catalysts. The first generation screen revealed α-hydroxy carboxylic acids and *N*-arenesulfonyl α-amino acids as promising ligand candidates. Based on these results, a focused second-generation array was constructed from the original 80 clusters to identify *N*-arenesulfonyl α-amino acids as superior ligands for the targeted reactions. Up to 76 % enantiomeric excess for the product was obtained with *N*-(toluene-4-sulfonyl)-L-leucine as ligand at − 78 °C (Figure 13).

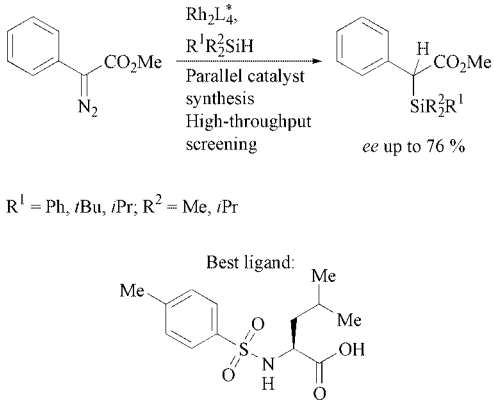

Figure 13. High-throughput screening of rhodium insertion catalysts for the carbenoid insertion of azo esters into Si–H bonds

In a related approach, pyridine-containing ligands could be indentified that, in the presence of copper, catalyze the Ullmann aryl ether reaction [73].

5. One-Pot, Multi-Substrate Screening

Hadamard transformation is a mathematical algorithm that can be used to identify the property of a particular candidate of a group by recording the sum or superpositions of signals from well-selected groups or pools of single wells, catalysts, etc. [74]. Subsequent deconvolution of a data set obtained from an entire group coupled with computer simulations can provide analytical solutions in various branches of physical sciences, including homogeneous and heterogeneous catalysis. Many applications and scenarios of Hadamard transformations in combinatorial catalysis can be envisioned including multisubstrate/one-catalyst screening, one-substrate/multicatalyst (multiple ligands, metals) screening, or simultaneous screening of a variety of catalysts against multiple substrates.

Consequently, Kagan et al. investigated a Hamadard-type approach to combinatorial catalysis [75]. They examined the performance of one catalyst against multiple substrates in the asymmetric reduction of mixtures (families) of

prochiral aromatic ketones to their corresponding alcohols. Corey's chiral oxazaborolidine catalyst was chosen as the chiral catalyst (Figure 14).

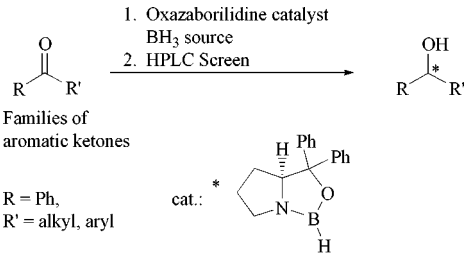

Figure 14. One-pot, multisubstrate screening of a chiral oxazaborolidine catalyst in the enantioselective reduction of a family of aromatic ketones

Conversion and enantiomeric excesses of the resulting mixture of alcohols were monitored by conventional HPLC on a chiral stationary phase (CSP). The key was the careful calibration of the screening conditions, i.e., baseline separation of all of the ketones (starting materials), the corresponding alcohols (products), and the enantiomers of the product alcohols. In addition, all ketone reductions were repeated individually in solution. The maximal discrepancy between the individual asymmetric reductions and the multisubstrate approach ranged from 1 to 7 %. For example, asymmetric reduction of acetophenone resulted in the formation of (*R*)-1-phenylethanol with 97 % *ee* under screening conditions (ketone mixture) and in the reduction of the isolated ketone.

The one-catalyst/multisubstrate strategy was also used for the pooled synthesis of libraries of internal olefins by metathesis [76], for the enantioselective addition of diethylzinc to aliphatic and aromatic aldehydes [77], and for the resolution of racemates [78].

6. Combinatorial Approaches to Olefin Polymerization Catalysts

An annual production of around 46×10^6 t exemplifies the industrial importance of polyolefins [79]. New and versatile catalysts with improved performance represent highly desirable research goals. Productivity, molecular weight capability, temperature stability, comonomer incorporation capability, controlled comonomer sequence distributions, and the copolymerization of new monomer combinations are sought to be optimized. Recent studies have uncovered catalyst systems based on middle and late transition metals that show polymerization activites comparable to those reported for commercially employed single-site metallocenes [80–82]. Of particular importance was a diimine-based palladium catalyst, which was reported to copolymerize ethylene with previously incompatible monomers such as methyl acrylate. These discoveries have led to considerable interest in new nonmetallocene catalysts because of their high activity, ease of synthesis, and in particular, their tolerance of functionalized olefins [83–86].

Combinatorial approaches can be employed for both the discovery and optimization of new catalysts and new polymers. The discovery process typically involves the identification of new catalyst systems that may be employed for improvements to existing processes, or novel processes for new polymers. Upon identification of a new catalyst system, a high-throughput optimization procedure can be used to further improve catalyst performance through a detailed examination of the influence of ligand structure, activators, scavengers, solvents, and temperature.

Powers, Murphy et al. and Symyx Technologies developed a parallel synthesis and screening protocol for a polystyrene-bound 96-member library of 1,2-diimine nickel and palladium complexes (Figure 15) [87, 88].

The key intermediate, a polystyrene-bound 1,2-diketone was converted to a 48-membered 1,2-diimine library through a condensation reaction with 48 electronically and sterically diverse anilines. The ligand library was then divided and converted into 48 nickel and palladium complexes by reaction with [$NiBr_2$(dme)] and [PdClMe(cod)]. An analogous solution-based 1,2-diimine library was also synthesized in order to compare the performances of the polystyrene-bound catalysts with those of the corresponding solution-based catalysts. A custom designed 48-cell high-pressure parallel polymerization reactor (PPR) was used to screen the catalyst library for ethylene polymerization activity (Figure 16) [89].

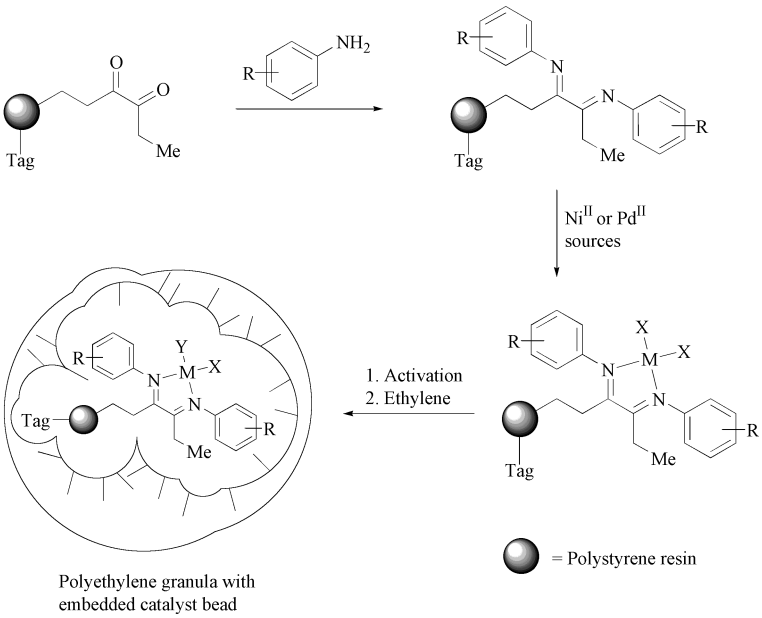

Figure 15. Synthesis of Brookhart-type polymer-bound Ni^{II} or Pd^{II} 1,2-diimine complexes and their use in the polymerization of ethylene
M = Ni^{II}, X = Y = Br; M = Pd^{II}, X = Me, Y = Cl; R = alkyl, aryl, heteroaryl, halogen, functional groups; tag = chemical code; MAO = methylalumoxane, Ar_F = 3,5-$(F_3C)_2C_6H_3$; dme = dimethoxyethane; cod = 1,5-cyclooctadiene.

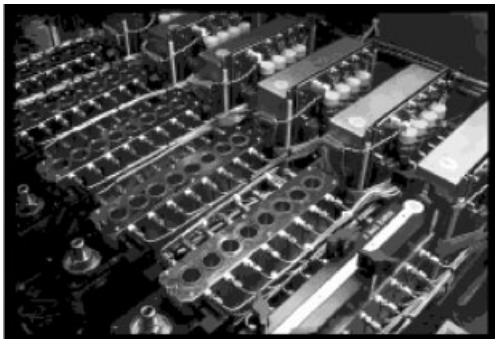

Figure 16. Parallel Polymerization Reactor-96 (PPR, Symyx Technologies)

Polymer products were assayed by a custom-designed rapid GPC technique [90]. The report also demonstrates the application of a chemical encoding strategy in which polystyrene-bound Ni^{II} and Pd^{II} complexes were encoded with cleavable tertiary amine tags. After a pooled polymerization, the tags were cleaved from the polymer products and detected by HPLC to reveal the identity of the catalyst.

The present work demonstrates the feasibility of applying solid-phase synthesis, on-bead screening, and encoding/decoding of pooled libraries of catalysts for the discovery and optimization of new olefin polymerization catalysts. More generally, Symyx has developed high-throughput methodologies for the synthesis and screening of supported and unsupported organometallic compounds for a range of important applications [91, 109].

Müllen et al. tagged silica- or polymer-supported heterogeneous catalysts for industrial olefin polymerization with a variety of fluorescent dyes [92]. Here, direct detection of the different product beads obtained by different catalysts is facilitated by fluorescent dyes that exhibit different emission wavelengths. The approach starts with producing the tagged catalysts by silica-supporting various metallocenes, activating them with MAO, and labeling them with different rylene dyes. The dyes were chosen because of their high chemical stability, their pronounced tendency to physisorb on silica, their high fluorescence yield, and because they do not influence the polymerization. These labeled catalysts were then mixed and introduced into a single polymerization vessel. During the olefin polymerization each catalyst par-

ticle forms only one product granulate through a particular growth process. To assign the different compounds of the product mixture to the employed catalyst, the polymer products are exposed to UV light, and can be directly identified, manually separated, and characterized, due to the different emission wavelengths of the labels. The feasibility of this concept was demonstrated for the polymerization of ethylene and for the copolymerization of ethylene and hexene.

Coates et al. used a pooled polymerization catalyst strategy to identify catalysts for the synthesis of syndiotactic polypropylene [93]. The basis of this concept is that the formed polymer itself serves as a stereochemical recording of the events on the polymerization catalyst. Assuming that the catalyst species do not react with one another, then a group of complexes for stereoselective polymerization can be screened simultaneously. When the desired polymer product has distinguishing chemical or physical properties, techniques such as solubility, spectroscopy, or chromatography can be used to quickly probe the crude product of a pooled polymerization reaction to see if a noteworthy catalyst is present. To probe this concept, a library of 12 salicylaldiminato ligands was synthesized separately by the condensation of three different salicylaldehydes and four amines (Figure 17). Equimolar amounts of these ligands were combined, deprotonated with BuLi, and treated with $TiCl_4$ to give a library of 78 possible titanium species. This complex library was then activated with MAO in toluene, and the resulting catalyst solution was exposed to propylene (ca. 0.3 MPa). Even though 90 % of the formed polymer could be washed away in refluxing diethyl ether (atactic polypropylene), the remaining 10 % of polymer was found to be syndiotactic polypropylene (^{13}C NMR microstructural analysis). Deconvolution methods, i.e., the synthesis and testing of sublibraries, were used to successfully identify the most active catalyst.

Chen et al. reported the use of electrospray ionization tandem mass spectometry (ESI-MS/MS) and gas-phase ion – molecule reactions for the rapid screening of Brookhart-type Pd^{II} olefin polymerization catalysts [94]. A test library of eight catalysts was prepared by treating eight diimine ligands with [PdClMe(cod)], washing and drying of the complex, and activation with AgOTf. The mixture was then dissolved in dichloromethane, saturated with ethylene, allowed to react for one hour and quenched with dimethyl sulfoxide (DMSO). After selection of ions with $m/z > 2200$ from the electrospray mass spectrum, these high-mass ions were collided with xenon to induce β-hydride elimination. From the corresponding daughter-ion spectrum the structure of the best catalyst can be derived.

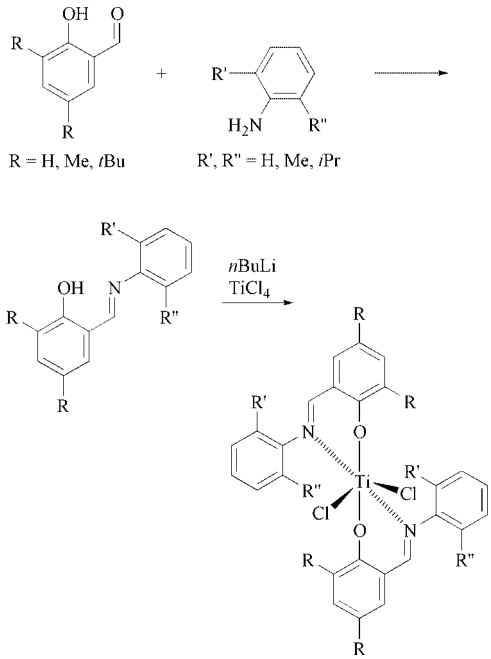

Figure 17. Synthesis of a library of salicylaldiminato titanium complexes

7. Combinatorial Inorganic Catalysis

The application of combinatorial approaches to inorganic homogeneous and heterogeneous catalysis has received a great deal of attention recently. This is largely because of difficulties encountered in exploring implicitly vast areas of phase or variable space and the empirical nature of conventional work that typifies research in inorganic catalysis. To date, a wide variety of catalytic processes has been studied by the combinatorial approach, ranging from polyoxometalates in oxidation catalysis, oxidative dehydrogenation of alkanes, CO oxidation, and NO reduction, to areas such as electrocatalysis,

fuel cell research, and electrosynthesis. High-throughput synthesis methods based on thin-film deposition, solution deposition, and impregnation have been developed, as well as exciting new tools for high-throughput characterization.

7.1. Combinatorial Libraries of Homogeneous Polyoxometalate-Based Catalysts

Early transition metal oxygen cluster anions or polyoxometalates (POMs) have a significant impact in a variety of different research fields. Among catalysis by heteropolyacids, selective hetero- and homogeneous oxidation processes are of utmost importance [95].

Hill and Gall reported the combinatorial synthesis of phosphorous-centered molybdenum and tungsten polyoxometalates of the Keggin type [96]. The POM catalysts were evaluated in the selective aerobic oxidation of tetrahydrothiophene (THT) to its corresponding sulfoxide (THTO; Figure 18).

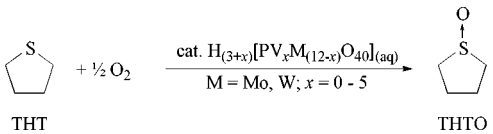

Figure 18. Aerobic oxidation of tetrahydrothiophene (THT) to tetrahydrothiophene sulfoxide (THTO) by combinatorially prepared POMs

The catalyst library was synthesized by combinatorially mixing stock solutions of appropriate tungsten, molydenum, and vanadium precursors at a constant phosphate concentration. Conventional gas chromatography was used to quantify product (THTO) yields and selectivities. Promising catalysts were selected and subsequently identified with the aid of ^{51}V NMR and IR spectroscopy, which revealed the complex $[\alpha\text{-}1,4\text{-}PV_2W_{10}O_{40}]^{5-}$ to be the most effective among them.

In a more recent contribution, Hill et al. reported a diversity-based discovery and mechanistic investigation of selective homogenous thioether oxidation by the $Au^{III}Cl_2NO_3/O_2$ system [97].

7.2. Combinatorial Libraries and High-Throughput Screening of Heterogenous Polyoxometalate Catalysts

Early efforts at Symyx Technologies led to the development of a fully integrated, high-throughput screening platform useful for pharmaceuticals and for fine-chemicals process discovery, screening, and optimization. The platform includes a proprietary software package for the design of combinatorial libraries. It also controls all the robotic functions, gathers data from each spatially addressable reaction entity, and automatically stores the process data in a central database. The hardware also comprises a variety of 96-well reactors equipped with a liquid- and slurry-handling robot. In one example, the integrated screening platform was employed for the high-throughput synthesis and screening of metal-doped polyoxometalate (POM) libraries in 96-well format [98]. Libraries of the heterogeneous POM catalysts were screened to discover selective aerobic oxidation catalyst systems. In this case, an imidazole alcohol was catalytically oxidized to the corresponding aldehyde, which represents an important key intermediate for the synthesis of DuPont's antihypertensive compound Losartan (Figure 19).

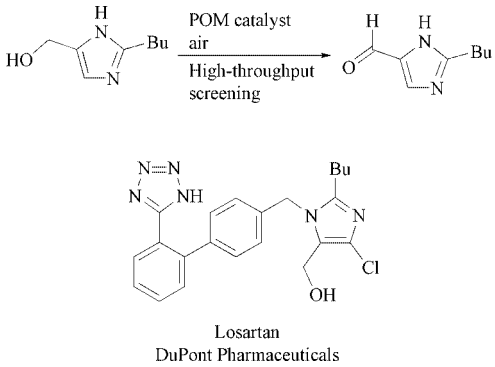

Figure 19. Application of high-throughput screening in the discovery of novel catalysts for the POM-catalyzed oxidation of imidazole alcohol to aldehyde, a key intermediate in the synthesis of the antihypertensive compound Losartan (DuPont Pharmaceuticals)

Within four weeks, 5000 experiments were performed to screen different catalyst compositions, effects of solvents, and additives under variable reaction conditions. Two classes

of catalyst systems that gave more than 90 % yield were identified. These results were confirmed and validated in subsequent scale-up experiments, but the proprietary catalyst composition was not disclosed.

8. Combinatorial Heterogeneous Catalysis

The mechanism of catalytic activity in a heterogeneous system is complex [99]. The distribution of active sites in heterogeneous catalysts and the phase integrity are in general poorly defined. The phase integrity plays a major role because the extent of deviation from the equilibrium structure under reaction conditions controls the catalytic function. These types of inherent problems severely hamper or even prevent the rational design of heterogeneous catalysts for a particular reaction. Conventional methods of catalyst discovery are primarily trial-and-error processes in which one catalyst is tested for activity at a time followed by numerous modifications to achieve satisfactory activity. The process therefore is time-consuming and laborious, and in most cases only local minima in a predefined parameter space (composition, reaction conditions, etc.) are identified rather than the absolute minimum. Combinatorial synthesis and screening offer a new dimension for the quick discovery and optimization of heterogeneous catalysts, both in terms of composition and process conditions. Integrated synthesis and screening of a plurality of catalysts in library format has been recognized as an essential factor [100]. For reviews on combinatorial heterogeneous catalysis, see [17, 19, 101, 102].

There are basically two techniques to prepare libraries of heterogeneous catalytic materials: solution-based methods and thin-film deposition methods (see Chap. 11). Impregnation of catalyst components onto a preformed solid support is an important solution-based technique for preparation of heterogeneous catalysts, but is seldom utilized for other materials science applications. The solution-based methods are predominant in the field of combinatorial heterogeneous catalysis. The majority of commercial heterogeneous catalysts are manufactured by solution-based techniques such as coprecipitation, impregnation, and variations thereof [103], and results from library experiments can thus be translated more easily into a bulk catalyst by using solution-based techniques. In addition, the solution-based methods present fewer scale-up problems. Furthermore, many liquid-handling robots and ink-jet-based liquid-dispensing systems can be employed to prepare heterogeneous catalyst libraries on a microgram to gram scale by solution-based methods. On the other hand, thin-film deposition methods are performed on custom-designed and -assembled high-vacuum instruments that generally require high initial investment. Special care should be taken to ascertain that the phase integrity of each catalyst member is the desired one, since the thin-film deposition methods rely on the interlayer diffusion for mixing, and the multicomponent layers occasionally result in phase separation. The thin-film method will perhaps be of choice for fused-catalyst libraries, e.g., metal alloys.

In most cases the catalyst performance depends on the method of preparation. Therefore, it is very important to adopt proper methods to obtain catalysts with expected surface area, uniform metal distribution, and desirable particle size. Newsam and Schüth have described various routes for combinatorial catalyst synthesis that include hydrothermal synthesis, use of fluid precursors, carrier impregnation, and precipitation [104]. Other methods of preparing combinatorial catalyst libraries include computer-controlled ink-jet deposition of liquid reactants [105], methods based on solution precursors [106–108], sol – gel techniques [109–111], impregnation of solid supports [112], and standard methods such as precipitation and coprecipitation [113, 114]. Table 1 summarizes some examples of heterogeneous catalyst libraries.

With the development of new and efficient technologies for the synthesis of large catalyst libraries, there is an ongoing effort for inventing fast and parallel screening tools for identifying active and selective catalysts. These techniques include resonance-enhanced multiphoton ionization (REMPI), time-resolved and differential IR thermography, scanning mass spectrometry, and colorimetry. While IR thermography and fluorescence or colored-dye assays are established techniques, a few new methods have emerged for identifying active catalysts in an array. So far, no screening method can offer a general solution to the problem of fast screening of

Table 1. Heterogeneous catalyst libraries

Catalyst libraries	Preparation method	Target reaction	Screening method	Ref.
0.5 % Ag, Bi, Co, Cr, Cu, Er, Fe, Gd, Ir, Ni, Pd, Pt, Rh, Ti, V, Zn, on Al_2O_3 pellets	solution-based impregnation	$H_2 + O_2$	IR thermography	[115]
37, 1 – 10 % Co, Cr, Cu, Fe, Ir, Mn, Ni, Pd, Pt, Rh, Ru, V, Zn on Si, Ti oxides	solution-based sol – gel method	hyrogenation of 1-hexyne	IR thermography	[116]
100 $Na_2O - Al_2O_3 - SiO_2$ zeolites with Li and Cs	solution-based hydrothermal zeolite synthesis	*	*	[117]
37 $TiO_2 - Al_2O_3 - SiO_2 - TiO_2 - ZrO_2$ zeolites	solution-based hydrothermal zeolite synthesis	*	*	[118]
645 clusters of Pt, Ru, Os, Ir on carbon	solution-based coprecipitation	methanol reforming	fluorescence acid – base indicator	[119]
120 ternary thin film clusters of Pt, Pd, Rh, and Pd, Rh, Cu	thin-film deposition: sputtering	CO oxidation, CO + NO	scanning quadrupole mass spectrometry	[120]
66 ternary combinations of 1 % Pt, Pd, and In on -Al_2O_3	solution-based impregnation	cyclohexane dehydrogenation to benzene	REMPI	[112]
16 Au/Co_3O_4 and Au/TiO_2 powders	solution-based coprecipitation	CO oxidation	quadrupole MS	[121]
33 1 – 6 % Ag, Au, Bi, Co, In, Cr, Cu, Fe, Mo, Ni, Re, Rh, Sb, Ta, Te, V, Y on Si, Ti, Zr oxides	solution-based sol – gel method	propylene oxidation	spatially resolved analysis of an array of batch microreactors	[122]
50 ternary and quaternary oxides of Co, Cd, Fe, Ga, Ge, In, Mn, Mo, Ni, Nb, V, W, Zn	solution-based coprecipitation	ethane and propane dehydrogenation	gas chromatography	[123]
30 binary combinations of Na_2WO_4 and Mn on SiO_2; Au and In on ZrO_2, TiO_2, SiO_2, MgO, ZnO, Nd_2O_3, Y_2O_3, CeO_2, Mn_2O_3	solution-based impregnation, deposition, and precipitation	oxidative coupling of methane and CO oxidation	quadrupole MS	[124]
66 ternary combinations of oxides of Mo, V, and Nb	solution-based sol – gel deposition	ethane oxidative dehydrogenation	scanning quadrupole MS, photothermal deflection, gas chromatography	[125]
144 ternary combinations of oxides of V – Al – Nb, and Cr – Al – Nb	solution-based sol-gel deposition	ethane oxidative dehydrogenation	scanning quadrupole MS, photothermal deflection	[126]
66 ternary combinations of 1 % Pt, Pd, and In on Al_2O_3	solution-based impregnation	cyclohexane dehydrogenation to benzene	quadrupole mass spectrometry	[127]
45 3 – 5 element combinations of Pt, Pd, Rh, Ru, Au, Cu, Ag, and Mn on TiO_2 and -Fe_2O_3	solution-based impregnation	C_3H_8 total oxidation	quadrupole mass spectrometry	[128]
56 quaternary combinations of Pt, Pd, In, Na on Al_2O_3	solution-based impregnation	NO reduction by C_3H_6	quadrupole mass spectrometry	[129]
$V_2O_5 - TiO_2$ mixtures	solution-based coprecipitation	oxidation on naphthalene to naphthoquinone	laser induced fluorescence imaging (LIFI)	[130]
36 Pt/Zr/V/Al_2O_3	solution-basedimpregnation	methane oxidation	scanning quadrupole MS coupled with monolith multi-channel reactor	[131]
52 Pt/Zr/V/Al_2O_3	solution-based impregnation onto Al_2O_3electrochemically formed on aluminum plate	methane oxidation	scanning quadrupole MS coupled with 35 parallel microreactors	[131]
715 combinations of Pt, Ru, Os, Ir, Rh	solution-based deposition	O_2 reductionH$_2$O oxidation	fluorescence acid – base indicator	[132]
280 V_2O_5, MoO_3, MnO_2, Fe_2O_3, Ga_2O_3, La_2O_3, B_2O_3, MgO on Al_2O_3	solution-basedimpregnation	propane dehydrogenation	quadrupole MS	[133]

* These publications only describe the synthesis of the libraries.

libraries together with the complexity involved in parallel detection of the reactants, products, and side products. Therefore, it is important to develop screening technologies according to the detection requirements of the reactants, products, and side products of the reaction. Accordingly, in some cases, it is essential to have a combination of tools for a better characterization of reaction products. It is also important to characterize the time-dependent nature of the catalyst library for developing a practical catalyst. Some catalysts have a significant induction period before they become fully active, and most catalysts deactivate over time. These properties will vary according to the reaction conditions, too. Some parallel screening schemes offer the capability of monitoring the reactions over time. For example, a multichannel fixed-bed reactor coupled with parallel GC and an array microreactor equipped with REMPI allow library screening over a prolonged reaction time. Many techniques such as mass spectrometry, IR spectroscopy, and other optical detection schemes can be applied in both homogeneous and heterogeneous catalysis. Techniques and tools for high-throughput catalyst screening are described in Chapter 10.

8.1. Oxidative Dehydrogenation of Ethane

The development of efficient heterogeneous catalysts for the gas-phase oxidative dehydrogenation of light paraffins is of particular interest because of the economic benefits of using paraffins directly as raw materials for the production of important base chemicals [125]. The low-temperature oxidative dehydrogenation of ethane to ethylene has been a research topic of consistent interest after the report of catalytic activity in the Mo – V – Nb – O system below 300 °C in 1978 [134, 135]. The optimum compositional range for these catalysts was reported to be $Mo_{0.61-0.77}V_{0.31-0.19}Nb_{0.08-0.04}$ [134, 135]. Researchers at Symyx Technologies have reproduced and improving these catalyst systems using combinatorial methods [136, 137].

The strategy reported is based on a three-step approach: In the primary screen thousands of catalyst formulations are generated by automated solution-based methods as arrays of thick films on 3-inch wafer plates by dispensing volumes on a nano- to microliter scale. The films themselves were prepared from sol – gel precursors that were deposited by automated solution deposition methods. The typical layout of a ternary library is shown in Figure 20. In this case the wafer holds a ternary ABC library consisting of two identical triangular matrices with 66 elements. The three corners in this library correspond to 100 % A, 100 % B, and 100 % C, respectively. Thus, the compositional increments are 10 % per matrix element for this kind of library layout.

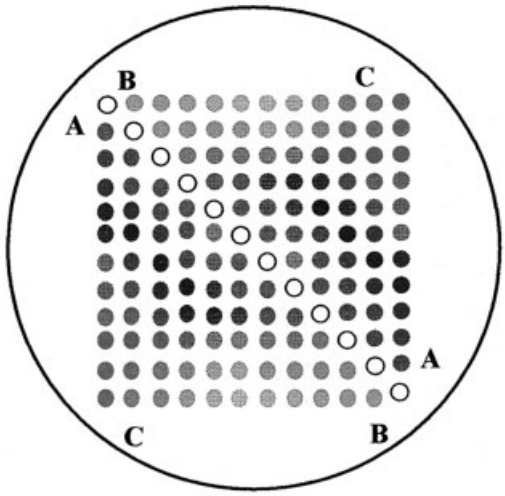

Figure 20. Library design for a 3-inch quartz wafer

The libraries were then screened by using simultaneous mass spectrometry and photothermal deflection spectroscopy. The best catalysts were consecutively synthesized on a 25 – 50 mg scale for the secondary screen in parallel fixed-bed reactors. The secondary screen provides conversion and selectivity data for the catalysts under realistic operating conditions. Catalysts that performed well in the secondary screen are further evaluated in tertiary screens in bench-scale reactors with gram quantities of catalysts.

To demonstrate the reliability of combinatorial methods compared to larger scale experimentation, Symyx scientists could show that the primary screening data with conversions in the ppm range correlate well with bulk catalyst activity with percentage-level conversions.

For better catalyst performance, researchers at Symyx Technologies investigated additional libraries focusing on the most active area. This fine-tuning of the catalyst resulted in an ethane conversion of 12.8 % and an ethylene selectivity of 74.4 % at an optimal composition of $Mo_{73}V_{24}Nb_3$. Including Sb, Ca, and Li as dopants improved the catalysts even further. Under identical synthesis and reaction conditions, the optimal composition reported by Thorsteinson, $Mo_{73}V_{18}Nb_9$, converts 6.1 % at a selectivity of 83.2 % [136, 137].

8.2. Oxidative Dehydrogenation of Propane

Recently, Baerns, Senkan et al. reported the impregnation method of catalyst preparation coupled with multitube packed-bed reactors [122] and array microchannel reactors [138] within the framework of genetic algorithms [139] to find catalysts for the low-temperature oxidation of propane at low concentrations [126]. The catalyst libraries were prepared by an automatic liquid-dispensation system applying the incipient wetness method. A total of eight active components were selected as candidates for incorporation into TiO_2 or Fe_2O_3 as carrier materials.

Different generations of libraries consisting of mixtures of active elements at different compositions were prepared and then tested for their catalytic performance (conversion of C_3H_8 to CO_2) in both multitube packed-bed reactors [123] and array channel microreactors [138]. Significant propane conversions could be achieved at 50 °C. Catalysts providing propane conversions on the order of 90 % were also achievable at higher temperatures. The best catalytic materials were found to be rich in ruthenium; multimetal combinations were significantly superior [126].

More recently, the same group disclosed an evolutionary approach to optimize multimetal oxide (V_2O_5, MoO_3, MnO_2, Fe_2O_3, Ga_2O_3, La_2O_3, B_2O_3, and MgO) catalysts supported on α-Al_2O_3 for the oxidative dehydrogenation of propane to propene [131]. Thus, 56 catalytic materials were prepared for each generation, and this process was iterated five times. Both manual synthesis and robotic synthesis were used for the library preparations, and these catalytic materials were tested in parallel. For the best materials propene yields of 7 % (1st generation) to 9 % (5th generation) were achieved.

8.3. Catalytic Oxidation of CO and Reduction of NO

These two processes were studied at Symyx Technologies using catalyst libraries of metal alloys of Rh, Pd, Pt, and Cu [119]. $15 \times 15 \times 15$ triangular libraries containing approximately 2 – 4 µg of 120 independently prepared catalysts were prepared by deposition with fully automated radio-frequency sputtering techniques. Libraries with similar catalysts were also prepared by sol – gel-based methods.

The oxidation of CO was carried out in the presence of oxygen, and the production of CO_2 was determined by scanning mass spectroscopy. The amount of CO_2 produced followed the trend Rh > Pd > Pt, which is in good agreement with results from experiments with single-crystalline surfaces and supported catalysts.

In subsequent libraries, one of the noble metals was replaced by transition metals. In the library with Rh – Pd – Cu, the 1:1 composition of Cu – Rh showed activity very similar to that of pure Rh, and the catalyst with 93 % Cu and 7 % Rh maintains 40 % of the activity of pure Rh. In addition, Pt – Cu and Pd – Cu binaries show activity enhancements over the respective pure metals, whereby the enhancement is stronger at low than at high temperatures.

The oxidation of CO by NO was investigated along the same lines with the above-mentioned Rh – Pd – Pt library. The reaction was carried out at ambient pressure at various temperatures. As in the oxidation of CO by oxygen, the activity for the oxidation of CO by NO follows the trend Rh > Pd > Pt, similar to that reported in the literature for bulk catalysts.

9. Combinatorial Electrocatalysis

Electrochemical reactions play important roles, from energy production in fuel cells and batteries to industrially important processes such as the production of chlorine and the electrohydrodimerization of acrylonitrile to adiponitrile

Figure 21. Photographs of the different stages of discovery of a Pt – Rh – Os ternary alloy in array format with a quinine indicator
A) Image in white light. B) Fluorescence immage at low overpotential, identifying the most active region of composition space. C) Fluorescence image at high overpotential, where methanol oxidation occurs at every spot in the array

used in the production of nylon. Electrocatalysis deals with the modification of the overall rates of electrochemical reactions so that selectivity of these reactions and their yield and efficiency are maximized. The main focus is on the electron transfer reactions that take place at the electrode – electrolyte interface, with electrochemical parameters such as potential and current being the major driving force for the reaction.

9.1. Electrocatalysts for Fuel Cells

Fuel cell technology holds the promise of low-to zero-emission energy for power plants, backup generators, and application in the transportation sector. Fuel cells also will offer fuel efficiency on the order of 50 % or more. With increasing interest in fuel cells as alternative energy sources for stationary power, portable, and automotive applications, there is a need to develop new anode and cathode electrocatalysts. The cost of the current precious metal based electrocatalysts is a considerable hurdle to the successful commercialization of fuel cells.

The oxidation of methanol in fuel cells has been extensively studied for several decades. At present, however, the commercialization of the direct methanol fuel cell (DMFC) is beset by several problems, e.g., the low loading levels of noble metal catalysts, the high cost of the platinum catalyst, and catalyst poisoning by CO.

Mallouk, Smotkin et al. developed a rapid combinatorial synthesis and fluorescence screening method for the exploration of ternary and quaternary anode electrocatalysts for the DMFC (Figure 21) [108]. The authors created a library of 135 different ternary alloys by using automated ink-jet deposition of metal precursor aliquots onto a conducting fiber paper, followed by reduction of the metal salts with a borohydride reagent. These compositions were based on five transition metals (Pt, Ru, Os, Rh, and Pd). After treatment the libraries were immersed in an aqueous methanolic sodium sulfate electrolyte containing a fluorescent pH indicator. Screening was performed with a typical three-electrode cell (electrode array as working electrode, Pt counterelectrode, and saturated calomel reference electrode) by conducting a single potential sweep from cathodic to anodic potentials. It is expected that the pH will drop close to the regions where methanol is being oxidized. The conducting Toray paper was illuminated with an UV lamp, and observation of the highest fluorescent intensities due to local proton release (pH drop) upon electro-oxidation of methanol determined the catalytically most active spots on a given library. The composition performances were compared to Pt/Ru (50/50 atom %). The best composition was given as Pt/Os/Rh (60/25/15 atom %).

In addition, the same authors prepared, screened, and analyzed a 645-member electrode array of Pt, Ru, Rh, Ir, and Os consisting of 80 binary, 280 ternary, and 280 quaternary combinations [140–142]. The most active catalyst was found to be Pt/Ru/Os/Ir (47/29/20/4 atom %).

Scientists at Symyx Technologies have developed another combinatorial approach to the rapid synthesis and screening of fuel cell electrocatalysts. Libraries of alloy materials are synthesized directly onto an 8×8 electrode array (Figure 22). The electrode array consists of 64 individually addressable electrodes prepared on an insulating 3-inch quartz substrate by lithographic techniques. The electrode can be made from gold or titanium. Each electrode is approx-

imately 1.0 mm indiameter, and electrical contact is made with each electrode via a contact pad on the edge of the wafer. Electrocatalysts were synthesized by combination of magnetron sputtering or by parallel electroplating. The electrode array was then sealed to a Teflon cylinder that was filled with aqueous electrolyte. All 64 electrodes were therefore exposed to the same electrolyte solution. A common reference electrode and working electrode were immersed in the cell. The array was interfaced to a 64-channel potentiostat via a PCB interface, and the electrocatalytic performance of the resulting materials were tested by parallel monitoring of the of the current – voltage – time behavior of each individual electrocatalyst.

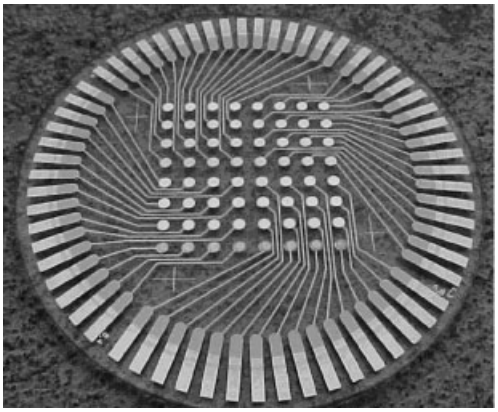

Figure 22. Symyx Technologies' electrode array consisting of 64 individually addressable electrodes prepared on an insulating 3-inch quartz substrate by lithographic techniques

The authors tested a number of catalytic concepts for the anodic electro-oxidation of methanol (DMFC), as well as the cathodic electroreduction of oxygen in aqueous acidic electrolytes [143]. It could be demonstrated that for known systems the electrochemical activity as measured directly on the thin-film samples on the addressable electrochemical array correlated with the activity of known powder samples.

9.2. Combinatorial Electrosynthesis

Ward et al. applied two different combinatorial approaches to study the electrochemical reduction of 1,4 benzoquinone to hydroquinone at organosulfur-modified gold electrodes [144]. The authors prepared an array of physically but not electrically isolated gold electrodes on a glass substrate. Monolayers of organosulfur reagents were prepared on selected electrodes on the 4×7 array by dispensing aliquots of hexanethiol or hexadecanethiol dissolved in ethanol at the respective electrodes. The array was subsequently rinsed and housed in a thin-layer electrochemical cell.

The first screening approach was based on a pH fluorescent assay, which was not sensitive enough. In a second approach the authors created an electrochemically addressable electrode array from which it was possible to directly measure the electrochemical activity by monitoring the current – voltage behavior at each individual electrode. In a serial manner the electrochemical reduction of benzoquinone was studied at each electrode by cyclic volatmmetry in which the potential of each electrode was cycled and the current measured. The authors proposed that this direct measurement of the electrochemical activity was more sensitive than the fluorescence screen in quantifying the benzoquinone reduction at the modified electrodes, and the results indicated that the activity in the library increases in the order of hexadecanethiol $<$ dodecanethiol $<$ hexanethiol $<$ gold, which indicates suppression of current with increasing alkanethiol chain length.

Yudin et al. demonstrated how combinatorial electrochemistry could be used in the electrosynthesis of small organic molecules [145]. The authors developed what they term a spatially addressable electrolysis platform (SAEP). Each electrochemical cell in the 4×4 array was equipped with a stainless steel cathode and a graphite rod anode. The cathodes were welded directly onto a stainless steel support that provided a common terminal for the current source. The authors explored the anodic oxidation of carbamates, amides, and sulfonamides leading to libraries of α-alkoxycarbamates, α-alkoxyamides and α-alkoxysulfonamides, respectively. Also, the intramolecular cyclization of hydroxyamides yielding heterobicyclic compounds and the generation of vicinal diamines by the reductive hydrocoupling of aldimines were achieved.

Yudin et al. generated libraries of catalytic materials on electrode surfaces by the copolymerization of bithiophene and pyrrole con-

taining TEMPO (2,2,6,6-tetramethilpiperidin-1-yloxy) catalysts [146, 147]. Diversity was created by electrochemical copolymerization and creating surfaces with different ratios of bithiophene/pyrrole. These catalyst films were utilized in the electrochemical oxidation of primary alcohols to aldehydes, whereby cyclic voltammetry was used to screen the catalytic activities of the modified electrodes.

Hillier et al. applied a scanning electrochemical microscope to characterize the oxidation of hydrogen on a polycrystalline platinum surface [148, 149] This technique could be utilized for the direct measurement of the rate constant for hydrogen oxidation and the reactivity mapping of heterogeneous electrodes consisting of catalytic and noncatalytic domains.

10. High-throughput Screening Tools for Catalysis

The process of discovering new catalysts or materials from a large pool of potential candidates requires a reliable and robust screening process. Traditionally, this process represents a bottleneck in the discovery effort, especially if large libraries of potential catalysts or materials have already been synthesized or prepared. The first step in combinatorial catalysis, like all materials design, involves the identification of a specific chemical transformation of interest. A collection of potential catalysts, prepared from a set of chemically diverse ligand sets and metal precursors, are combined in a parallel or combinatorial fashion and screened for activity in a high-throughput primary screen. Although the large volume of potential catalysts examined in a primary screen typically allows only relative activity to be established, the goal at this stage of discovery is to identify promising lead catalysts worthy of further investigation and follow up. Lead materials are then further examined in a high-throughput secondary screen designed to screen for specific trends in the physical and chemical properties in greater detail and typically at a slightly lower throughput than the primary screen. The information obtained from the secondary screen is then used to prepare additional generations of catalysts that can be optimized into a superior catalyst worthy of commercialization.

Typically, the information necessary to classify new materials cannot be obtained from a single characterizing device. Therefore, a series of high-throughput screening tools are employed at various stages in the combinatorial process. While a large number of automated commercial systems exist for high-throughput analysis of microliter samples for medical and pharmaceutical applications, the majority of the screening tools necessary for advanced materials research are custom-made instruments. Conventional analytical tools such as mass spectrometry, gas [150] and liquid chromatography, electrophoresis,[151–153], Raman [154] and NMR spectroscopy, X-ray fluorescence microprobe analysis [133], and X-ray diffraction [155] have been automated and redesigned for rapid serial measurements of hundreds of samples per day [156]. A number of reviews on high-throughput screening techniques are available [17, 18, 20, 21, 102, 157–159]. Several examples of novel high-throughput screening systems are listed in the following:

1) Infrared
   a) IR thermography [34, 115, 116, 160–166]
   b) High-throughput infrared spectroscopy [167–170]
2) Optical
   a) Colorimetric assay [108, 132, 171–179]
   b) Resonance-enhanced multiphoton ionization (REMPI) [112, 130, 138]
   c) Photothermal deflection [125]
   d) Circular dichroism [180, 181]
3) Mass spectrometry
   a) Scanning MS [120, 122, 182, 183]
   b) Electrospray MS using mass tags [184, 185]
4) Electronic methods
   a) Electrode arrays [119, 144, 145, 147, 148, 186]
   b) Solid-state temperature sensors [187]
   c) Solid-state gas sensors [188]
5) Eelectrophoresis
   a) Capillary array electrophoresis [151–153]

11. Combinatorial Solid-State Materials Science

The application of combinatorial technologies towards discovery in solid-state materials sci-

ence has exploded during the past few years. The central challenges faced in adaptation of high-throughput techniques are related to the ability to prepare solid-state materials in an array format while retaining key physical properties. To this end, significant technological advances have been made in the areas of the synthesis and characterization of solid-state materials by adaptation of conventional solution and vapor-phase synthetic techniques. A wide variety of high-throughput characterization methods for arrays of solid-state materials are now available, including optical and structural techniques such as electro-, cathodo-, and photoluminescence, as well as X-ray diffraction. Using these techniques, significant advances continue to be made in the areas of superconductivity, ferromagnetic superconductors, magnetoresistant materials, dielectric and ferroelectric materials, and various luminescent materials.

11.1. Materials Library Synthesis

One of the biggest challenges posed by the high-throughput synthesis of solid-state materials in an array format is the need to maintain relevant bulk physical properties. To this end, a variety of synthesis techniques based on solution deposition or thin-film deposition have been developed for the preparation of libraries of solid-state materials [189–194]. The choice of synthesis technique is based on the material being prepared as well as the primary screen employed after synthesis. For example, by modifying technologies similar to those used to make integrated circuit (IC) chips, materials libraries or integrated materials (IM) chips were first developed and utilized by Schultz et al. [10]. Regardless of the synthetic technique utilized, a library of solid-state materials must typically undergo thermal processing involving temperatures in excess of 400 °C to complete the synthesis of the target materials. Such high temperatures required the development of thermally stable substrates that allow the preparation of two-dimensional, spatially addressable arrays of solid-state materials.

11.1.1. Vapor Deposition Techniques

Vapor deposition is commonly used in the semiconductor industry to deposit thin films of material onto a substrate. Vapor deposition techniques that have been utilized in combinatorial library synthesis include sputtering [21, 195, 196] and thermal evaporation [197–201], electron-beam evaporation [202, 203], pulsed laser ablation [21, 204, 205], ion-beam implantation [206–209], molecular-beam epitaxy [210–213], and chemical vapor deposition [214, 241].

One of the most straightforward thin-film approaches is the continuous-composition spread (CCS) technique, which utilizes two or three off-axis sources to co-deposit material on a substrate [196]. This technique relies on the non-uniform deposition of materials formed by the geometric arrangement between the sources and the substrate. The relative concentration of each component at a specific location on the substrate decreases with the distance from the source. As materials spread from the sources, they mix in the vapor and are deposited on the substrate with atomic-level mixing, which reduces or eliminates the need for high-temperature post-processing of the library. This technique is also amenable to the isolation of metastable or low-temperature phases that are crystalline on deposition. The lack of precise stoichiometric control and limited compositional range has relegated this technique primarily to optimization and exploration of systems with only two independent variables.

Another technique frequently employed for combinatorial materials science utilizes a vacuum deposition system with several source materials, used in conjunction with masking techniques employing physical or shadow masks, movable shutters, or photolithography [10, 189, 190, 215–218, 253]. Masking allows for controlled deposition of different materials sequentially or simultaneously on particular areas of the substrate. The design of the masks and the sequence in which they are employed determine how the materials are deposited at a given location on the substrate. By altering the sequence, time, and rate of deposition, it is possible to control the exact chemical composition of each element in the library.

Figure 23 illustrates the efficiency of such a masking approach, whereby a compositionally variant array of materials can be constructed by using physical or photolithographic masks. In

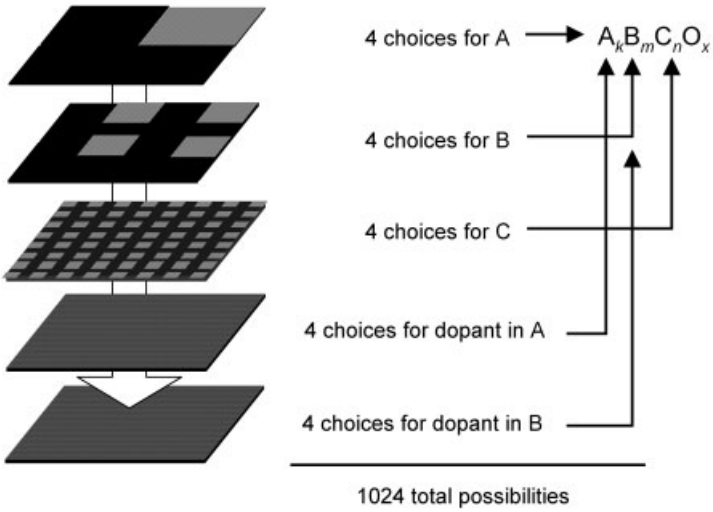

Figure 23. The quaternary masking system uses six fractal masks, each of which can be rotated by 90° to allow four choices of materials at each level

this example, a portion of the substrate is alternatively covered with a mask and subjected to vapor deposition of a particular component of $A_kB_mC_nO_x$ until each component has been deposited. Gradient deposition is often employed to allow systematic control over variations in the stoichiometry for each component and the thickness of the depositions. The geometry and positional orientation of the masks can be varied in many ways, and this allows the phase space to be systematically and exponentially constructed. For example, in the hypothetical phase space of $A_kB_mC_nO_x$ in Figure 23 the use of four different components for A, B, and C utilizing four dopant components for A and B results in 4^5 or 1024 unique compositions.

Thin-film deposition methods are synthetically quite versatile; they have progressed to enable atomic- and molecular-layer epitaxy and offer the ability to construct artificial lattices, epitaxial overlayers, and patterned films of a variety of materials. Dopants are usually sandwiched between layers of the host material to avoid evaporation and to assure proper interdiffusion. Subsequent thermal processing is usually required to ensure the formation of single-phase materials in the library of materials or devices [189, 219, 220, 250]. Physical properties of the materials can be subsequently screened with contact or noncontact screening probes. The number of compounds that can be simultaneously synthesized by this technique is limited by the spatial resolution of the masks (up to 10^6 per square inch if photolithography is used) and by the degree to which microscale synthesis can be performed.

Combinatorial thin-film techniques such as those described above have been utilized in the discovery of a variety of new electronic materials. For example, Pessaud et al. utilized pulsed laser ablation of two compound targets made of superconducting $YBa_2Cu_3O_7$ (YBCO) and the double-chain insulator $MCuO_2$ (M = Ca, Sr) to create new metastable superconducting materials [204].

New multiwavelength-emitting chips for applications in wavelength-divisible multiplexing (WDM) have also been explored. For example, Liu et al. succeeded in growing single quantum wells of $Al_{0.35}Ga_{0.65}As$/GaAs on GaAs (001) substrates by MBE and were able to generate more than 20 different wavelengths from a single chip [207].

Amorphous materials have been investigated in a combinatorial optimization of hydrogenated amorphous silicon (a-Si:H) based thin-film transitor (TFT) devices by Koinuma et al. using plasma-enhanced chemical vapor deposition [214].

11.1.2. Alternative Library Synthesis Techniques

The synthesis of many materials such as zeolites, polymers, polycrystalline phosphors, and many other oxides and sulfides is best accomplished by using solution-phase methods. When applicable, solution-phase methods allow the creation of a large number of libraries at a significantly lower capital investment compared to automated vapor deposition equipment. Additionally, solution-phase methods can often result in improved compositional control over thin-film techniques. One significant complication associated with materials discovery in thin-film format is the occasional lack of correlation with bulk properties due to differences caused by film microstructure, strain, and so on. Physical and chemical properties of materials prepared by solution techniques generally show excellent correlation with bulk properties, regardless of the quantity of material prepared. This is particularly advantageous from the standpoint of enabling the correlation of properties of materials synthesized in a library format with those prepared in bulk by traditional solid-state synthesis.

A major advantage of solution techniques is that they allow mixing at the molecular level, which reduces the need for high-temperature interdiffusion and also facilitates the isolation of metastable phases. For example, sol – gel techniques are frequently employed in the preparation of arrays of a variety of inorganic nanostructured materials such as phosphors [221], oxide semiconductors [222], and polymer/vanadium oxide nanocomposites [223]. The use of soluble metal precursors such as metal alkoxides in the fabrication of such arrays makes the sol – gel approach particularly amenable to combinatorial methods [221, 222].

Solution-based techniques can also be use to prepare highly dense arrays of materials. For example, Xiang, Schultz et al. have demonstrated that a scanning multihead ink-jet delivery system can be used to perform automated microsynthesis of solid-state libraries. Using this technique, deposition volumes on the order of 500 nL can be dispensed rapidly with an accuracy better than 99 % [190, 224]. The system has been successfully used to generate libraries with 100 members per square inch, and a system capable of generating 1000 members per square inch is under development.

The application of solution-based techniques in the hydrothermal synthesis of materials such as multicomponent zeolites has also been explored. To this end, a variety of reactors for hydrothermal synthesis have been designed to accommodate the requirements of small-scale synthesis [225–227]. For example, Akporiaye et al. have reported the design of stackable, miniature, multi-autoclave reactors that house up to 100 reaction wells with volumes of ≤ 0.5 mL [225]. Using a battery of these reactors, the ternary phase diagram for the $Na_2O - Al_2O_3 - SiO_2 - H_2O$ system was successfully reproduced with total sample volumes ≤ 0.5 mL. Klein et al. have reported the construction and testing of an even smaller reactor, that employs a silicon substrate for hydrothermal synthesis and can achieve reaction volumes on the order of 2 μL. Libraries of materials can then be analyzed with an automated X-ray diffraction instrument [226]. The success of these and other approaches clearly suggest that general microscale synthesis of materials by hydrothermal synthesis is tenable.

The adaptability of solution-based combinatorial strategies towards the preparation of nanoclusters and surfaces exhibiting nanoscale morphologies has also been demonstrated. For example, Baker et al. have described the synthesis of surfaces exhibiting nanoscale variation in mixed-metal compositions and architecture [228]. Continuous or stepped gradients in the size and number density of surface features can be generated simultaneously over different regions of a single substrate. For example, a library of Ag-clad colloidal Au arrays was prepared in which the coverage of colloidal Au and the extent of the reductively deposited Ag coating were varied in an attempt to optimize the enhancement factor for surface enhanced Raman scattering (SERS) [228, 229]. In one application of nanoscale-based synthesis, Wright et al. have shown that combinatorially derived peptide and peptidomimetic ligands based on phytochelatins can stabilize CdS nanoclusters [230–232].

Lithographic masking techniques analoguos to those described for the preparation of thin-film arrays of materials have been used in conjunction with solution-based approaches. For example, Vossmeyer et al. have developed a scheme for generating complex, spatially sep-

arated patterns of multiple types of semiconducting and/or metallic nanocrystals employing photolithography. The technique involves the stepwise preparation of multiple-particle arrays based on lithographic patterning of amino-functionalized organic monolayers that contain a photolabile protecting group covalently bound to a substrate with an SiO_2 surface. The substrate is photolyzed, which results in deprotection of the unmasked areas, and then treated with amine-stabilized nanoparticles of metals that bind to the exposed SiO_2 surface. This technique was used iteratively to prepare binary micropatterns consisting of Pt or Au and highly luminescent CdSe/CdS core/shell nanocrystals [194].

Many new advanced materials ranging from magnetic materials to bulk metallic glasses involve alloying multiple elements to control both intrinsic and extrinsic properties. Permanent magnets, for example, may contain a dozen or more elements to control grain boundaries and domain wall movement. Cohen-Adad et al. utilized a melting-zone technique to investigate AlCo and AlCoFe alloys [233]. In their method, source materials are cut into wedges and then welded together into a bar. The bar is then placed in an alumina crucible and drawn through a horizontal zone furnace to facilitate mixing. The resulting bar is diced into pieces, and the composition at each point along its length is measured by microprobe analysis. The morphology and crystal structure of the materials were investigated with optical microscopy, differential thermal analysis, and X-ray diffraction. Inoue has reviewed similar techniques used to search for new glasses, which typically contain five or more elements to inhibit crystallization [234].

11.2. High-Throughput Screening for Materials

The large numbers of compounds and the speed with which they may be prepared by combinatorial synthesis require commensurate automated high-throughput characterization. Typically, multiple methods of characterization are required to adequately screen new materials during primary, secondary, and higher order screening. To this end, a series of high-throughput screening tools are often employed at various stages in the combinatorial process to create a screening hierarchy. The screening tools required for high-throughput characterization are frequently custom-made instruments, often incorporating or adapting technologies developed for conventional materials characterization. Conventional analytical tools such as mass spectrometry, gas and liquid chromatography, optical and NMR spectroscopy, X-ray fluorescence microprobe analysis [235], and X-ray diffraction have been automated and redesigned for rapid serial measurements of hundreds of samples per day [156, 236].

11.2.1. Optical Screening

Optical screening methods, where applicable, offer high-throughput noncontact screening for combinatorial libraries. Optical screens typically involve a laser or fiber-optic probe that is serially scanned across a library, or utilize cameras capable of imaging library elements in parallel. In studies of luminescent materials, high-throughput screens using UV-excited photoluminescence allow quantitative measurement of chromaticity and relative extrinsic efficiencies [110, 111]. Optical transmission and reflection measurements have also been utilized to determine the band gap of various oxide thin films, as well as to estimate band gaps for thermoelectric and photovoltaic materials [213, 237]. Lastly, an apparatus for the measurement of electro-optic coefficients of thin-film samples that functions by measuring the polarization modulation of light from a HeNe laser passing through a sample has been described by Xiang. This technique can be used to determine electro-optic coefficients of materials in a thin-film format [10].

11.2.2. X-Ray Characterization

Structural characterization by X-ray diffraction plays a critical role in conventional synthesis of solid-state materials and is equally important in high-throughput methods. Several groups have created instruments capable of X-ray microdiffraction on samples in array format. [235, 236]. This has been made possible by the use of intense X-ray sources (rotating anode or synchrotron), and focusing mirrors that collimate an X-ray

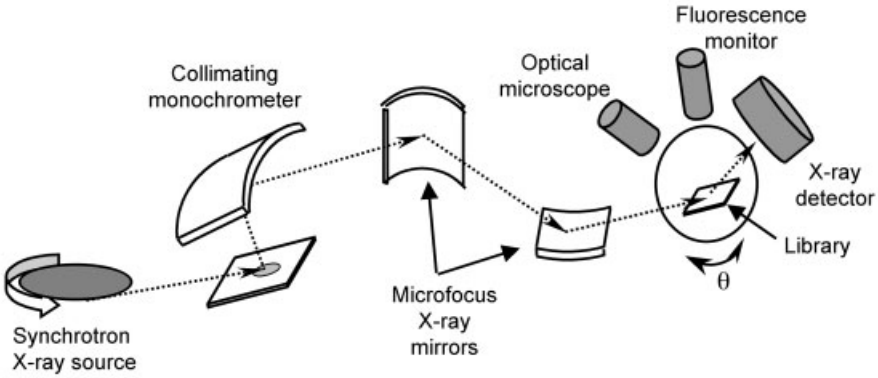

Figure 24. Schematic representation of an X-ray microprobe synchroton beamline capable of producing a $3 \times 20\,\mu m^2$ X-ray spot on a sample in the library array

beam onto a small spot within the dimensions of the sample size in the array (Figure 24). However, data collection times for the measurement of a significant number of library elements can require hours (rather than minutes). Moreover, exploratory libraries are typically mixed phases in nature, whose complicated analysis has yet to be dealt with efficiently.

11.3. Applications of Materials

11.3.1. Superconductivity

One of the first demonstrations of combinatorial thin-film materials exploration by Schulz et al. involved the preparation of libraries of superconducting copper oxide thin films by sputtering several metal oxides in sequence onto an inert single-crystal support of either MgO or $LaAlO_3$ [10, 238, 190]. Arrays containing different combinations, stoichiometries, and deposition sequences of $BaCO_3$, Bi_2O_3, CaO, CuO, PbO, $SrCO_3$, and Y_2O_3 were generated. The superconductivity of each sample was evaluated by measuring its resistance as a function of temperature, with an array of probes that analyze 64 samples at a time. Superconducting films of $BiSrCaCuO_x$, $BiPbCaSrCuO_x$, and $YBa_2Cu_3O_x$ were identified with critical temperatures in the range of 80 – 90 K.

Xiang et al have developed a scanning evanescent microwave microscope (SEMM) capable of nondestructive and quantitative mapping of electrical impedance with submicrometer resolution [241, 242]. Unlike conventional four-probe methods, the SEMM is a noncontact method, which eliminates dependence on contact resistance and interfacial effects. The microscope is routinely used for characterization of linear and nonlinear dielectric constants of ferroelectrics and dielectrics.

Alternative sample pooling is used in high-T_C superconductor research [243, 244]. Superconductivity is one of the few electronic properties that can be measured when samples are pooled (mixed together) in a single vial. In this method, several compositions are synthesized separately and then mixed together to form a single sample. The presence of a Meissner fraction indicates the presence of superconductivity in at least one of the compounds in the mix, the identity of which must be determined by individually measuring each of the components in that particular mix.

Bozovic et al. utilized a sophisticated composition-spread MBE system, an ozone source, an electron microscope (LEEM/LEED/PEEM system), time-of-flight ion scattering and recoil spectroscopy, and scanning quartz microbalances in a search for new materials for oxide electronics [245].

11.3.2. Ferromagnetic Semiconductors

Ferromagnetic semiconductors are important materials for spin-based electronics in which the correlation between charge and spin is used

to generate spin-dependent electronic functionality such as giant magnetoresistance and spin field effect transistors. Among known materials, Mn-doped GaAs is the most promising for practical applications because it has the highest magnetic ordering (Curie) temperature ($T_C \approx 100$ K. The search for new ferromagnetic semiconductors has led to the investigation of dilute magnetic oxides. Kionuma et al. have applied a combinatorial approach to the search for new ferromagnetic semiconductors by adding small amounts of cobalt to the anatase form of TiO_2 [246]. TiO_2 and $Ti_{0.5}Co_{0.5}O_2$ ceramic targets were ablated with a KrF excimer laser in a combinatorial laser molecular beam epitaxy system [247]. The ferromagnetic ordering temperature was estimated to be in excess of 400 K, a substantial increase over the Mn-doped GaAs compounds.

11.3.3. Magnetoresistant Materials

The preparative combination of thin-film deposition and physical masking techniques was used by Xiang et al. in the discovery of new cobalt oxide magnetoresistance materials [109, 248, 249]. Magnetoresistance is the variation of electrical resistance with an applied magnetic field. Materials with particularly large magnetoresistance properties are useful in high-density magnetic storage technology such as read heads for computer hard drives [250]. Colossal magnetoresistance (CMR) was first discovered in Mn-based perovskites $(La,R)_{1-x}A_xMnO_{3-\varepsilon}$, where R is a rare earth element, and A is Ca, Sr, or Ba. Two identical 128-member libraries of composition $Ln_xM_yCoO_z$ were synthesized on $LaAlO_3$ single crystals, where Ln = $(La, Y)^{3+}$ and M = $(Ca, Sr, Ba, Pb)^{2+}$. The two libraries were annealed and sintered under different conditions, and one showed an overall decrease in magnetoresistivity, which highlights the fact that in combinatorial materials synthesis, post-deposition processing conditions represent important variables. Large magnetoresistance (Δ; $R/R(0) > 5$ %) was measured in $La_x(Ba,Sr,Ca)_yCoO_z$ samples, while Y-based samples exhibited much smaller magnetoresistive effects.

11.3.4. Dielectric and Ferroelectric Materials

Driven by the current trend toward the continued miniaturization of electronic devices, improved thin-film insulators are needed in, for example, dynamic random access memory (DRAM) computer chips. Using the composition spread technique (CCS), van Dover et al. from Lucent Technologies have reported new dielectric thin-film materials that have higher dielectric coefficients than amorphous silicon dioxide (a-SiO_x), the insulating material most commonly used today [242, 251–253]. The CCS technique involves the co-deposition of three different reagents onto a substrate from a triangular arrangement of targets. Van Dover et al. used off-axis RF magnetron sputtering guns to codeposit a continuous ternary composition variation of Zr, Ti, and Sn on a TiN-coated Si wafer [251, 253]. A continuous thin film was obtained, and the position-related composition was inferred by using Rutherford backscattering spectroscopy. To isolate the optimum thin-film transistor candidate ($Zr_{0.15}Sn_{0.3}Ti_{0.55}O_{2-\varepsilon}$), more than thirty combinatorial libraries were synthesized and evaluated at a rate of approximately one per day.

Combinatorial approaches have been applied by Xiang et al. to ferroelectric materials in which the effects of transition metal dopants on the dielectric coefficient and tangent loss of a library of thin films of $(Ba_xSr_{1-x})TiO_3$ (BST) were determined [254, 255]. Ferroelectric materials are being widely used in the development of new microwave devices such as frequency agile filters, phase shifters, and tunable resonators, and $(Ba_xSr_{1-x})TiO_3$ compounds have been intensely studied. Four libraries of 256 differently doped thin films of $(Ba_xSr_{1-x})TiO_3$, where $x = 1, 0.8, 0.7$, and 0.5, were generated on a $LaAlO_3$ substrate. These four hosts were then doped with different combinations of different metallic elements. The resulting dopant map for the $(Ba_{0.8}Sr_{0.2})TiO_3$ quadrant is shown in Figure 25 (see next page).

Appropriate post-annealing afforded high quality epitaxial thin films, comparable to those made by conventional in situ methods from stoichiometric targets on hot substrates. The rapid, quantitative and nondestructive characterization of the microwave dielectric coefficient and loss tangent of samples in the library was achieved

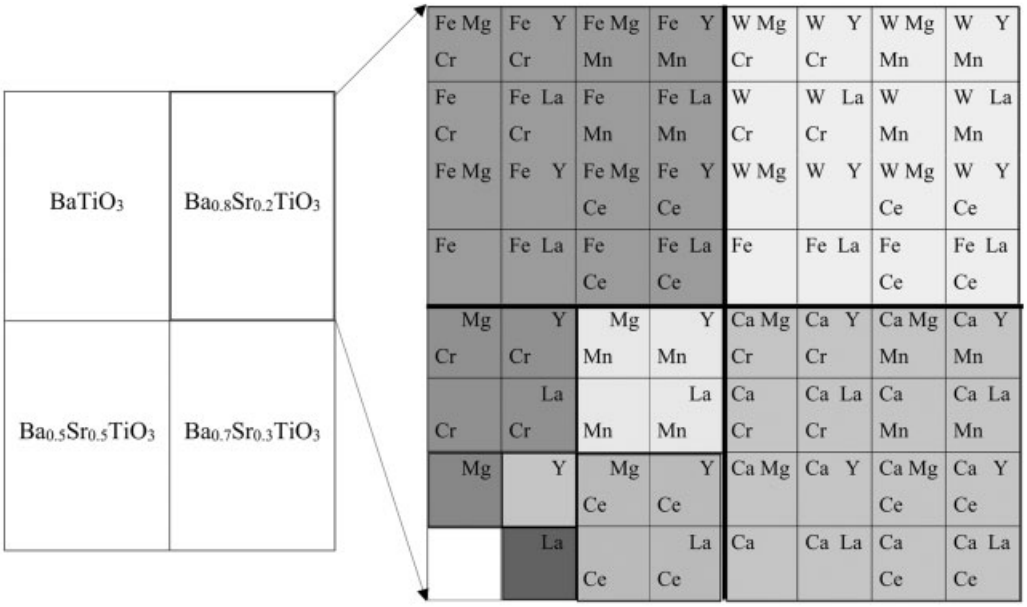

Figure 25. Dopant map for the $(Ba_{0.8}Sr_{0.2})TiO_3$ quadrant of a ferroelectric thin-film library; the same map applies to the other three quadrants containing different host materials. A quaternary masking scheme with four masks was used to deposit 256 differently doped thin films of $Ba_xSr_{1-x}TiO_3$ $(0.5 < x < 1)$ on a 2.54×2.54 cm $LaAlO_3$ substrate.

with a scanning-tip microwave near-field microscope at 1 GHz, and the results were found to be consistent with measurements made with interdigital electrodes [21, 239, 256–261].

Subsequently, the effects of dopants on device performance were evaluated by incorporating a layer of the doped dielectric $(Ba_xSr_{1-x})TiO_3$ in a parallel-plate capacitor library between Pt and $La_{0.5}Sr_{0.5}CoO_3$ electrodes [262]. Multilayer device structure libraries of 240 different host/dopant combinations were epitaxially synthesized with sharp interfaces from amorphous layers on a $LaAlO_3$ substrate. Since the conventional techniques of high-temperature, in situ epitaxial thin-film growth cannot be used for combinatorial library synthesis, Xiang et al. developed a synthesis method to form epitaxial multilayers from amorphous layers deposited at room temperature using pulsed laser deposition (PLD).

11.3.5. Luminescent Materials

The combinatorial methodology of thin-film deposition and masking techniques to create materials chip libraries has quickly found applications in the discovery and optimization of luminescent materials. Despite many years of intensive research, fewer than 100 useful commercial phosphor materials have been discovered by conventional one-at-a-time synthesis and testing. Phosphor materials with high quantum efficiencies find application in display technologies such as cathode ray tubes (CRTs) and are of great importance for developing improved plasma, field emission, and electroluminescent flat-panel displays, and lighting [263]. Oxides are attractive host materials for the development of advanced phosphors due to their ease of synthesis and stability. For example, the efficient refractive oxide phosphor, $Y_2O_3{:}Eu^{3+}$ (red), $Y_3Al_5O_{12}{:}Tb^{3+}$ (green), and $BaMgAl_{10}O_{17}{:}Eu^{2+}$ (blue) have found applications in tricolor lamps, projection TVs, and plasma displays.

Symyx Technologies' red phosphor $Y_{0.845}Al_{0.07}La_{0.06}VO_4{:}Eu_{0.025}$, with a quantum efficiency approaching those of existing commercial red phosphors, was discovered in

a library of 25 000 different compositions produced by using electron-beam evaporation to deposit thin films onto a 3-inch silicon wafer [110]. In this discovery library, four constant-thickness columns were first deposited, consisting of SnO_2, V_2O_5, $Al_2O_3 + V_2O_5$, or Al_2O_3. On top of these layers were deposited four rows of linearly varying thickness of La_2O_3, Y_2O_3, MgO, and $SrCO_3$ to create 16 host lattice subregions. Finally, within each of the 16 subregions linearly varying thickness of the rare earths Eu_2O_3, Tb_4O_7, Tm_2O_3, and CeO_2 were deposited, resulting in approximately 600 different chemical compositions per square centimeter (Figure 26). High-throughput screening for UV-excited photoluminescence was performed by photographing the visible emission of the library with a charge-coupled device (CCD) camera while exciting with a 254 nm broadband source. To optimize the best host compositions of the Eu-doped red phosphors, $Y_{1-m}Al_mVO_4$, identified in the initial high-density exploration library, a second library was deposited on a triangular grid, which included La as a third host component to explore all possible $Y_{0.95-m-n}Al_nLa_mVO_4$:$Eu_{0.05}$ compounds. The maximum intensity with red chromaticity corresponds to the phosphor composition $Y_{0.82}Al_{0.07}La_{0.06}VO_4$:$Eu_{0.05}$. The activator concentration in this lead host was subsequently optimized with a third library in which the Eu^{3+} concentration was varied, resulting in the final optimized composition $Y_{0.845}Al_{0.07}La_{0.06}VO_4$:$Eu_{0.025}$ with improved red chromaticity ($x = 0.67$, $y = 0.32$) compared to the more orange standard commercial red phosphor $Y_{1.95}O_3$:$Eu_{0.05}$ ($x = 0.64$, $y = 0.35$). Synthesis of bulk samples by conventional methods confirmed the performance of the newly identified phosphor composition.

The group from Symyx Technologies followed the red vanadate phosphor with a report on a novel luminescent oxide, Sr_2CeO_4 with a bluish white emmision [111].

Xiang et al. prepared several 128-member libraries of aluminates of the formula $Gd(La,Sr)AlO_x$ doped with rare earth activators. The combinatorial libraries were used to survey the phosphorescence of a wide range of oxide compositions under a variety of processing conditions [264].

Subsequently, Xiang et al. demonstrated that a scanning multi-inkjet delivery system can be used to achieve results identical to the vapor-deposited libraries [265].

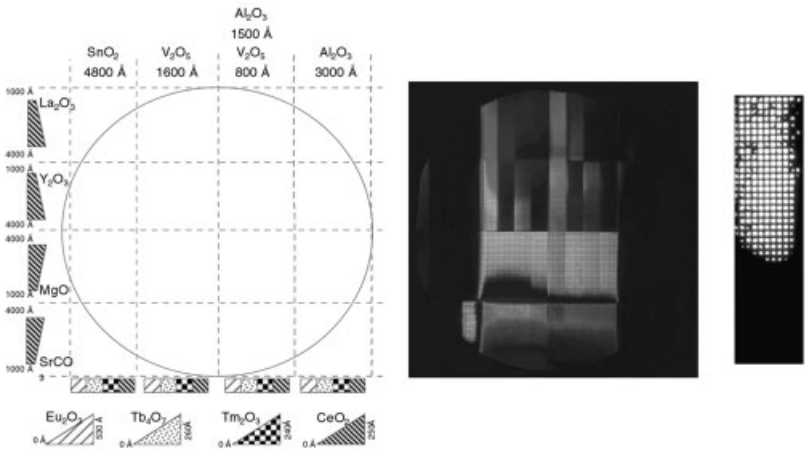

Figure 26. Deposition map (left) of a diverse phosphor discovery library; the thicknesses of the various materials deposited on a 3-inch silicon wafer are shown. Photograph (right) of the UV-excited photoluminescence of the discovery library containing approximately 25 000 different compositions. In this library, novel efficient red and blue phosphor compositions were identified and subsequently optimized in focus libraries.

Libraries of thin-film phosphors $(Gd,La,Y)_m(Ta,Zr,W,Mo,Zn)_n(Al,Mg,Sr)_kO_x$:$(Tm,Eu,Tb,Ce)_y$ (with $0 \leq m$, n, $k \leq 1$, $0.005 \leq y \leq 0.1$) were generated by Sun and Xiang using sequential RF [266]. A new red phosphor with the composition $(Gd_{1.54}Zn_{0.46})O_{3-\delta}:Eu^{3+}_{0.06}$ was claimed.

Schultz et al. used RF/DC sputtering to prepare three 1024-member libraries of silicate and gallate host materials on Si substrates, each library entity containing a number of different dopants [218].

Thin-layer organic light-emitting devices (OLEDs) are an emerging technology that is beginning to be commercialized in flat panel displays. To develop this technology further, it is important to optimize different materials properties, layer thicknesses, deposition parameters, and layer compositions under comparable and reproducible conditions. A combinatorial approach has been developed to create libraries of different OLEDs in order to screen and optimize device parameters with respect to power efficiency, photometric efficiency, and current – voltage characteristics [197–200, 267]. Schmitz et al. used a combinatorial approach to optimize the Alq_3 (8-hydroxyquinolinlato – Al^{III}) layer thickness on top of a constant-thickness hole-transport layer (HTL) to resolve whether an additional electron-transport layer (ETL) can further enhance device performance [200].

12. Organic Materials and Polymers

The application of combinatorial chemistry holds great promise in the discovery of novel functional organic materials. The current emphasis is on finding materials that either selectively bind small organic molecules or have improved physical and mechanical properties.

12.1. Schiff Bases for Nonlinear Optical (NLO) Materials

An investigation of the combinatorial synthesis procedures needed for the rapid formation of large libraries of polar compounds for the preparation of nonlinear optical (NLO) materials has been described by Nesterov et al. [268]. The structural characteristics of the Schiff base products (i.e., donor and acceptor groups connected to a π-conjugated chain) are believed to have potential as NLO or electro-optical materials. Nesterov et al. created a test library using four amines and two aldehydes as the building blocks for eight potential polar Schiff bases. Acid-catalyzed condensation of a set of electron-rich cinnamaldehydes with four electron-poor anilines generated a small library of eight Schiff bases. Crystalline precipitates were observed in every reaction, and crystals suitable for X-ray diffraction were obtained from all but one of the reactions. Five crystal structures were determined, all with centrosymmetric space groups that preclude the compounds from being NLO candidates.

12.2. Artificial Receptors for Small Organic Molecules

Since the rational design of nanoporous materials and their inclusion cavities based on molecular structures is difficult, combinatorial methods were implemented to find superior lattice inclusion compounds [269]. When screened against a guest molecule (*i*PrOH), seventeen ammonium carboxylate host compounds were discovered from a combinatorial library of 100 salts formed by mixing ten commercially available amines and ten carboxylic acids. One host compound was identified that forms inclusion compounds with 25 different organic solvents. Application to chiral recognition by the proper choice of amines and carboxylic acids to form a variety of host compounds is under investigation.

Leblanc et al. applied combinatorial principles to surface chemistry by synthesizing and screening libraries of amphiphilic lipid molecules [270]. At the air – water interface, the hydrophilic polar moiety of these molecules is embedded into the water phase and the hydrophobic alkyl chain is oriented towards the air phase. When compressed at the interface, the amphiphilic molecules move and assemble into organized supramolecular structures, known as Langmuir monolayers. The polar moiety was functionalized with a peptide library such that the self-assembly of these functional lipids formed proteinlike supramolecular structures. One library with 250 members and three

sublibraries of 54 members each were assayed for molecular recognition of maltose and other sugar molecules. One sublibrary, containing both polar charged (Glu) and aromatic (Tyr) amino acid residues, was identified that selectively bound maltose but not D-glucose and sucrose.

12.3. New Materials for the Separation of Enantiomers

Given the enormous demand for enantiomerically pure compounds in the pharmaceutical field, novel chiral stationary phases (CSPs) for high-performance liquid chromatography (HPLC) were developed by combinatorial means [271]. Fréchet et al. prepared a model library of chiral stationary phases (selectors) for HPLC by attaching a small model library of 36 chiral compounds that were obtained from three L-amino acids and 12 aromatic amines to synthetic polymer beads (Figure 27) [272]. After the entire on-bead library of selectors was packed into a HPLC column, selectivity was assayed by injecting various racemates of chiral targets. Despite the presence of 36 mixed selectors, racemic DL-(3,5-dinitrobenzoyl)leucine diallylamide was separated. A deconvolution process was used to determine which of the 36 selectors in the HPLC column was the most powerful. In contrast to the "one-column, one-selector" approach, which would require the preparation and testing of 36 chiral stationary phases, this combinatorial scheme requires the preparation of only 11 columns for the discovery of the most selective material. With the same approach, a library of 240 selectors could be deconvoluted with only 17 columns.

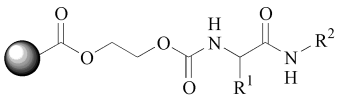

Selector: R^1 = alkyl, R^2 = aryl

= poly(2-hydroxyethyl methacrylate-*co*-ethylene dimethacrylate)

Figure 27. Model library of chiral stationary phases (selectors) for HPLC

Two groups applied the principle of reciprocity to the identification of new CSPs [273–275]. This approach is based on the following assumption: if a ligand of the stationary phase is selective toward an analyte, then the same analyte immobilized on the proper support should result in a stationary phase that is very likely to be selective toward the original ligand [276]. Following this principle, the chiral substrate (*S*)-(3,5-dinitrobenzoyl)leucine acted as an "inverse" selector and was immobilized on synthetic polymer beads [273]. A library of 108 racemic 4-aryl-1,4-dihydropyrimidines (DHPM) was screened against the "inverse" selector by observing the enantioselectivity for resolution. The best candidate was then attached to monodisperse macroporous beads to prepare the new polymer-based chiral stationary phase which proved efficient in the resolution of several analytes.

A different approach is to synthesize parallel combinatorial CSP libraries with up to 200 members and screen them for their ability to resolve a racemate of the analyte [277–279]. These libraries of potential selectors typically consist of immobilized dipeptides and a capping group. Racemic analyte is then added to each member of the library and allowed to equilibrate. The enantiomeric ratio of the analyte in the supernatant is analyzed after the equilibration period by circular dichroism measurements or other methods. A selective absorption of one of the two enantiomers to the resin is indicative of a chiral selector.

Welch et al. introduced a method for rapid LC/MS screening of chiral stationary phases based on the use of isotopically labeled enantiomers [280]. "Pseudoracemates" made up of a pair of isotopically differentiated pseudoenantiomers were prepared and screened against different CSPs. The enantioenrichment of the supernatants can be rapidly estimated by comparing the MS abundance of the ions corresponding to each of the pseudoenantiomers.

A visual screen for identifying chiral selectors was introduced by Still et al. [281]. The idea was to treat an equimolar mixture of a blue L-amino acid and a red D-amino acid with a library of chiral selectors on synthesis beads, each of which carries a different selector (Figure 28). Highly enantioselective binding results in red or blue beads, whereas unspecific binding yields brown beads. The reddest and the bluest beads found in the assay were picked and decoded to

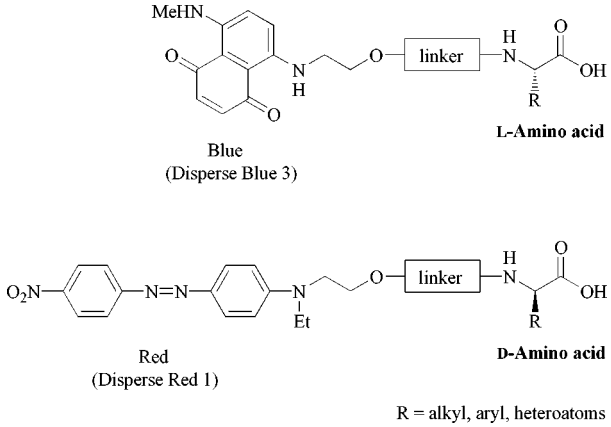

Figure 28. Enantiomeric probe labeled with different dyes

determine the structure of the bound chiral selector.

12.4. Molecular Imprinting

Molecularly imprinted polymers (MIPs) are another class of artificial receptors for small organic molecules. Their binding performances give these materials great potential in combinatorial approaches as recognition matrices for the screening and rapid selection of ligands from a combinatorial library [282–284]. In molecular imprinting recognition sites are generated within a macromolecular matrix by use of a molecular template in a casting procedure [285]. The selected ligand (print molecule) is first allowed to interact freely in solution by bond formation with one or more functional monomers. The resulting adducts are subsequently copolymerized with a large excess of cross-linker to give a rigid, insoluble polymer. Following extraction of the print molecule, specific recognition sites are left in the polymer. The technique has been applied to the screening of a combinatorial steroid library [282], and a combinatorial library of molecularly imprinted polymers has been tested in chiral separations [284].

In 1999 Takeuchi et al. reported a rapid combinatorial approach to synthesize and evaluate artificial receptors [286]. The preparation of the MIPs was automatically performed by using programmable liquid-handling equipment with a new in situ molecular imprinting protocol. The freshly synthesized MIP library (polymer films) was then subjected to a primary screening to estimate the affinity of the template to the resultant polymer. A secondary screening on purified MIPs then allows for a more precise evaluation of the affinities and selectivities of the new materials. MIPs that are artificial receptors for triazine herbicides ametryn and atrazine were prepared by the combinatorial molecular imprinting approach by using variable amounts of two functional monomers, methacrylic acid (MAA) and 2-(trifluoromethyl)acrylic acid (TFMAA), as well as ethylene glycol dimethacrylate as a cross-linker and AIBN as the polymerization initiator.

12.5. Polymer Topologies and Functionalization

Combinatorial approaches have also been applied to polymer chemistry. For example, a small library of 112 strictly alternating copolymers (0.2 g per sample; Figure 29 [see next page], bottom) was prepared from 14 distinct tyrosine-derived diphenols and eight aliphatic diacids [287]. Up to 32 polymers were synthesized in parallel under identical reaction conditions, and conventional methods were used to characterize the polymers providing structure – property correlations.

In a different approach, polymerization reactions between different combinations of two starting materials have been found to lead to uniquely responding sensors with responses not

simply related to the proportions of the starting materials [288].

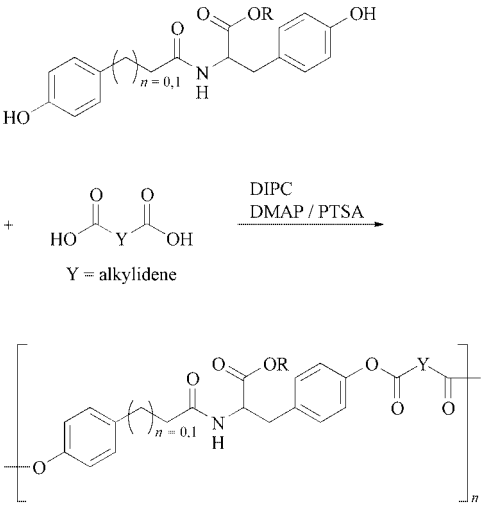

Figure 29. Library of polyarylates derived from diphenols and diacids

Radical polymerizations have been the focus of a variety of combinatorial approaches. Janda et al. applied a sequence of normal and living free-radical polymerizations to generate a library of 20 block copolymers possessing either block or graft architecture [289]. The members of the block copolymer library were then screened for their ability to serve as a soluble polymer support for organic reactions. One block copolymer could be identified that has a solubility profile that is complementary to the current soluble polymer of choice in liquid-phase organic synthesis (LPOS), namely polyethylene glycol (PEG).

Hawker et al. used combinatorial methods to identify novel nitroxide-based mediators for living free-radical polymerizations [290, 291]. Since the structure of the nitroxide moiety plays an important role in the success of living free-radical polymerizations, a structurally diverse, α-hydrogen-bearing nitroxide library was prepared and screened for the ability to control the polymerization of a variety of functionalized monomers. One α-hydrido alkoyamine was identified as the most promising candidate to effect the living free-radical polymerizations of a wide variety of monomers, e.g., styrenes, acrylates, acrylamides, and acrylonitrile-based monomers.

Researchers at Symyx Technologies have focused on combining rapid synthesis and rapid characterization techniques with full automation to reap the maximum benefit of combinatorial chemistry in the area of polymer science [292, 293]. Utilizing the stanard microtiter plate format, a 96-well, glass-lined aluminum reactor was developed to conduct organic reactions above the boiling point of the reagents without loss of pressure. Two methods of rapid molecular mass determination, i.e., precipitation – redissolution chromatography (PRC) and rapid-fire light scattering (RFLS), were used to screen the resulting polystyrene samples.

Combinatorial approaches were also employed to produce π-conjugated polymers with novel electrical and/or photofuntional properties. The photoluminescence of π-conjugated polymer thin films makes them interesting materials for flat panel displays. Since film preparation parameters are numerous, combinatorial concepts have been successfully applied for the synthesis of π-conjugated polymer thin films [294]. A variety of π-conjugated polymer building blocks was used for this study (Figure 30). Combinatorial film preparation was achieved by using appropriate mask shapes and synchronizing the rotation of the substrate with the target exchange. Vacuum evaporation (VE) or pulsed laser deposition (PLD) was used to deposite films on silica and/or Si substrates.

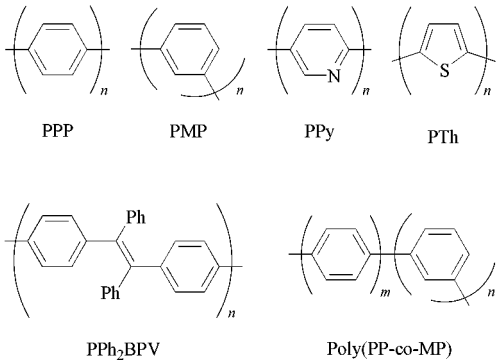

Figure 30. π-Conjugated polymers utilized for the combinatorial fabrication of thin films: PPP: poly(*para*-phenylene), PMP: poly(*meta*-phenylene), PPy: poly(pyridine-2,5-diyl), PTh: poly(thiophene-2,5-diyl), PPh2BPV: poly(diphenylene biphenylene vinylene), Poly(PP-co-PM): poly (*para*-phenylene-co-*meta*-phenylene)

13. Summary and Outlook

Combinatorial approaches to discovery and optimization of catalysts and materials will inevitably become the paradigm for modern scientific research. Considerable headway has already been made in this direction, as can been seen from the tremendous breadth of applications and approaches discussed herein. A multitude of tools are now available using combinatorial technologies that appropriately accommodate the new tasks and requirements for accelerated preparation of new materials. A common underlying theme associated with these technologies is miniaturization, parallelization, and automation so that large numbers of samples can be synthesized and screened efficiently. Rapid serial and parallelized adaptations of conventional analytical techniques will become increasingly important in the assay of materials properties, as will the development and implementation of new and unconventional high-throughput screening tools. Software development and engineering support in the construction and design of synthesis and screening tools are as crucial as further advances in chemistry, even when appropriate tools or robots for synthesis and screening automation are commercially available. Finally, the combinatorial methodology generates data much faster than the conventional research employing "empirical" and "rational" approaches to materials discovery, and, inevitably, the proper data handling and storage should accompany the high-throughput synthesis and screening to maintain the integrity of research and development efforts. Full realization of the combinatorial methodology will require the integration of chemistry, physics, engineering, and informatics to greatly enhance the probability of finding a global reaction optimum of a catalyst for a targeted reaction or materials with desired properties. In the future, the scientist's intuition may be shifted, at least to some degree, towards optimally programming and setting up appropriate experiments and screens as well as for the analysis of the obtained data and materials. All these efforts will require an enormous initial investment in all of these areas.

14. References

Specific References

1. G. Jung (ed.): *Combinatorial Chemistry - Synthesis, Analysis, Screening,* Wiley-VCH, Weinheim, 1999.
2. E. M. Gordon, J. F. Kerwin (eds.): *Combinatorial Chemistry and Molecular Diversity in Drug Discovery,* John Wiley & Sons, New York, 1998.
3. N. K. Terret: *Combinatorial Chemistry,* Oxford University Press, Oxford, 1998.
4. S. R. Wilson, A. W. Czarnik (eds.): *Combinatorial Chemistry,* John Wiley & Sons, New York, 1997.
5. B. A. Bunin (ed.): *The Combinatorial Index,* Academic Press, San Diego, 1998.
6. I. M. Chaiken, K. D. Janda (eds.): *Molecular Diversity and Combinatorial Chemistry,* American Chemical Society, Washington, 1996.
7. R. Dagani, *Chem. & Eng. News* **7710** (1999) 51 – 60.
8. M. A. Gallop, R. W. Barrett, W. J. Dower, S. P. A. Fodor, E. M. Gordon, *J. Med. Chem.* **37** (1994) 1233 – 1251.
9. For Chemoinformatics see: M. Hann, R. Green, *Curr. Opin. Chem. Biol.* **3** (1999) 379 – 383.
10. X.-D. Xiang, X.-D. Sun, G. Briceno, Y. Lou, K. A. Wang, H. Chang, W. G. Wallace-Freedman, S.-W. Chen, P. G. Schultz, *Science* **268** (1995) 1738 – 1740.
11. J. J. Hanack, *J. Mater. Sci.* **5** (1970) 964 – 971.
12. F. Zaragoza Dörwald (ed.): *Organic Synthesis on Solid Phase,* Wiley-VCH, Weinheim, 2000.
13. S. Booth, P. H. H. Hermkens, H. C. J. Ottenheijm, D. Rees, *Tetrahedron* **54** (1998) 15385 – 15443.
14. R. Hanko, K. C. Nicolaou (eds.): *Combinatorial Chemistry,* Wiley-VCH, Weinheim, 2002, in press.
15. For accounts and special techniques used in combinatorial chemistry, see for example: A. W. Czarnik, J. A. Ellman (eds.): Special Issue on Combinatorial Chemistry, *Acc. Chem. Res.* **29** (1996) 111 – 170.
16. M. J. Cannarsa, T. Uno, C. Larsen, *Curr. Opin. Drug Discov. Dev.* **3** (2000) 743 – 749.
17. S. Senkan, *Angew. Chem.* **113** (2001) 322 – 341; *Angew. Chem. Int. Ed.* **40** (2001) 312 – 329.
18. M. T. Reetz, *Angew. Chem.* **113** (2001) 292 – 320; *Angew. Chem. Int. Ed.* **40** (2001) 285 – 310.

19. F. Gennari, P. Seneci, S. Miertus, *Catal. Rev.-Sci.* **42** (2000) 385 – 402.
20. B. Jandeleit, D. J. Schaefer, T. S. Powers, H. W. Turner, W. H. Weinberg, *Angew. Chem.* **111** (1999) 2648 – 2689; *Angew. Chem. Int. Ed.* **38** (1999) 2495 – 2532.
21. D. R. Liu, P. G. Schultz, *Angew. Chem.* **111** (1999) 36 – 56; *Angew. Chem. Int. Ed.* **38** (1999) 36 – 54.
22. For a comprehensive overview see: B. Cornils, W. A. Herrmann: *Applied Homogeneous Catalysis with Organometallic Compounds,* Wiley-VCH, Weinheim, 2000.
23. D. Maclean, J. J. Baldwin, V. T. Ivanov, Y. Kato, A. Shaw, P. Schneider, E. M. Gordon, *J. Comb. Chem.* **2** (2000) 562 – 578.
24. For a review on enzyme mimics see: A. J. Kirby, *Angew. Chem.* **108** (1996) 770 – 790; *Angew. Chem. Int. Ed.* **35** (1996) 705 – 724.
25. F. M. Menger, Z. X. Fei, *Angew. Chem.* **106** (1994) 329 – 331; *Angew. Chem. Int. Ed.* **33** (1994) 346 – 348.
26. F. M. Menger, A. V. Eliseev, V. A. Migulin, *J. Org. Chem.* **60** (1995) 6666 – 6667.
27. P. D. Miller, W. T. Ford, *Chem. Commun.* 1998, 1151 – 1152.
28. F. M. Menger, C. A. West, *J. Ding, Chem. Commun.* 1997, 633 – 634.
29. F. M. Menger, J. Ding, V. Barragan, *J. Org. Chem.* **63** (1998) 7578 – 7579.
30. T. Berg, A. M. Vandersteen, K. D. Janda, *Bioorg. Med. Chem. Lett.* **8** (1998) 1221 – 1224.
31. T. Berg, A. Simeonov, K. D. Janda, *J. Comb. Chem.* **1** (1999) 96 – 100.
32. A. Berkessel, D. A. Hérault, *Angew. Chem.* **111** (1999) 99 – 102; *Angew. Chem. Int. Ed.* **38** (1999) 102 – 105.
33. M. Havranek, A. Singh, D. Sames, *J. Am. Chem. Soc.* **121** (1999) 8965 – 8966.
34. S. J. Taylor, J. P. Morken, *Science* **280** (1998) 267 – 270.
35. I. Ojima (ed.): *Catalytic Asymmetric Synthesis,* VCH, Weinheim, 2000.
36. For a review see: K. Soai, S. Niwa, *Chem. Rev.* **92** (1992) 833 – 856.
37. G. Liu, J. A. Ellman, *J. Org. Chem.* **60** (1995) 7712 – 7713.
38. For reviews on polymer-supported reagents see: O. Brümmer, B. Clapham, K. D. Janda, *Curr. Opin. Drug Discov. Devel.* **3** (2000) 462 – 473.
39. For a review on polymer-assisted purification techniques see: J. J. Parlow, R. V. Devraj, M. S. South, *Curr. Opin. Chem. Biol.* **3** (1999) 320 – 336.
40. For a review on polymeric scavenger reagents see: J. Eames, M. Watkinson, *Eur. J. Org. Chem.* 2001, 1213 – 1224.
41. I. Chataigner, C. Gennari, U. Piarulli, S. Ceccarelli, *Angew. Chem.* **112** (2000) 953 – 956; *Angew. Chem. Int. Ed.* **39** (2000) 916 – 918.
42. A. J. Brouwer, H. J. van der Linden, R. M. J. Liskamp, *J. Org. Chem.* **65** (2000) 1750 – 1757.
43. For a reviews on hetero Diels-Alder methodology see: K. A. Jorgensen, *Angew. Chem.* **112** (2000) 3702 – 3733; *Angew. Chem. Int. Ed.* **39** (2000) 3558 – 3588.
44. S. Kobayashi, K. Kusakabe, H. Ishitani, *Org. Lett.* **2** (2000) 1225 – 1227.
45. A. M. Porte, J. Reibenspiess, K. Burgess, *J. Am. Chem. Soc.* **120** (1998) 9810 – 9187.
46. For a review on transition metal catalyzed asymmetric allylic alkylation see: B. M. Trost, D. L. V. Vranken, *Chem. Rev.* **96** (1996) 395 – 422.
47. S. R. Gilbertson, C.-W. T. Chang, *J. Org. Chem.* **63** (1998) 8424 – 8431.
48. S. R. Gilbertson, S. E. Collibee, A. Agarkov, *J. Am. Chem. Soc.* **122** (2000) 6522 – 6523.
49. J. Matsuo, K. Odashima, S. Kobayashi, *Org. Lett.* **1** (1999) 345 – 347.
50. L. S. Hegedus: *Transition Metals in the Synthesis of Complex Organic Molecules,* University Science Books, Mill Valley, 1999.
51. D. G. Blackmond, *CatTech* **2** (1998) 17 – 31.
52. S. R. Gilbertson, X. Wang, *Tetrahedron* **55** (1999) 11609 – 11618.
53. B. M. Cole, K. D. Shimizu, C. A. Krueger, J. P. A. Harrity, M. L. Snapper, A. H. Hoveyda, *Angew. Chem.* **108** (1996) 1776 – 1779; *Angew. Chem. Int. Ed.* **35** (1996) 1668 – 1671.
54. M. S. Sigman, E. N. Jacobsen, *J. Am. Chem. Soc.* **110** (1998) 4901 – 4902.
55. K. D. Shimizu, B. M. Cole, C. A. Krueger, K. W. Kuntz, M. L. Snapper, A. H. Hoveyda, *Angew. Chem.* **109** (1997) 1782 – 1785; *Angew. Chem. Int. Ed.* **36** (1997) 1704 – 1707.
56. C. A. Krueger, K. W. Kuntz, C. D. Dzierba, W. G. Wirschun, J. D. Gleason, M. L. Snapper, A. H. Hoveyda, *J. Am. Chem. Soc.* **121** (1999) 4284 – 4285.
57. J. R. Porter, W. G. Wirschun, K. W. Kuntz, M. L. Snapper, A. H. Hoveyda, *J. Am. Chem. Soc.* **122** (2000) 2657 – 2658.
58. S. J. Degrado, H. Mizutani, A. H. Hoveyda, *J. Am. Chem. Soc.* **123** (2001) 755 – 756.

59. J. R. Porter, J. F. Traverse, A. H. Hoveyda, M. L. Snapper, *J. Am. Chem. Soc.* **123** (2001) 984 – 985.
60. C. A. Luchaco-Cullis, H. Mizutani, K. E. Murphy, A. H. Hoveyda, *Angew. Chem.* **113** (2001) 1504 – 1508; *Angew. Chem. Int. Ed.* **40** (2001) 1456 – 1460.
61. M. B. Francis, E. N. Jacobsen, *Angew. Chem.* **111** (1999) 987 – 991; *Angew. Chem. Int. Ed.* **38** (1999) 937 – 941.
62. Y. Hoshino, H. Yamamoto, *J. Am. Chem. Soc.* **122** (2000) 10452 – 10453.
63. K. Burgess, H.-J. Lim, A. M. Porte, G. A. Sulikowski, *Angew. Chem.* **108** (1996) 192 – 194; *Angew. Chem. Int. Ed.* **35** (1996) 220 – 222.
64. M. T. Powell, A. M. Porte, K. Burgess, *Chem. Commun.* 1998, 2161 – 2162.
65. S. J. Taylor, J. P. Morken, *J. Am. Chem. Soc.* **121** (1999) 12202 – 12203.
66. M. Kawatsura, J. F. Hartwig, *Organometallics* **20** (2001) 1960 – 1964.
67. S. Bromidge, P. C. Wilson, A. Whiting, *Tetrahedron Lett.* **39** (1998) 8905 – 8908.
68. C. Moreau, C. G. Frost, B. Murrer, *Tetrahedron Lett.* **40** (1999) 5617 – 5620.
69. J. Christoffers, A. Mann, *Angew. Chem.* **112** (2000) 2871 – 2874; *Angew. Chem. Int. Ed.* **39** (2000) 2752 – 2754.
70. M. A. Huffmann, P. J. Reider, *Tetrahedron Lett.* **40** (1999) 831 – 834.
71. R. T. Buck, D. M. Coe, M. J. Drysdale, C. J. Moody, N. D. Pearson, *Tetrahedron Lett.* **39** (1998) 7181 – 7184.
72. Available Chemical Directory ACD: MDL, Information Systems, Inc., San Leandro, 14600 Catalina Street, CA 94577, USA, www.mdli.com.
73. P. J. Fagan, E. Hauptman, R. Shapiro, A. Casalnuovo, *J. Am. Chem. Soc.* **122** (2000) 5043 – 5051.
74. A. Marshall (ed.): *Fourier, Hadamard and Hilbert transforms in chemistry,* Plenum Press Corp., New York, 1982.
75. X. Gao, H. B. Kagan, *Chirality* **10** (1998) 120 – 124.
76. C. Brändli, T. R. Ward, *Helv. Chim. Acta* **81** (1998) 1616 – 1621.
77. C. Gennari, S. Ceccarelli, U. Piarulli, C. A. G. N. Montalbetti, R. F. W. Jackson, *J. Org. Chem.* **63** (1998) 5312 – 5313.
78. T. Vries, H. Wynberg, E. van Echten, J. Koek, W. ten Hoeve, R. M. Kellogg, Q. B. Broxterman, A. Minaard, B. Kaptein, S. van der Sluis, L. Hulshof, J. Kooistra, *Angew. Chem.* **110** (1998) 2491 – 2496; *Angew. Int. Ed.* **37** (1998) 2349 – 2353.
79. N. Kashiwa, J. Imuta, *Cat. Surv. Jpn.* **1** (1997) 125 – 142.
80. B. L. Small, M. Brookhart, A. M. Bennett, *J. Am. Chem. Soc.* **120** (1998) 4049 – 4050.
81. For a related class of olefin polymerization catalysts see: G. J. P. Britovsek, V. C. Gibson, B. S. Kimberley, P. J. Maddox, S. J. McTavish, G. A. Solan, A. J. P. White, D. J. Williams, *Chem. Commun.* 1998, 849 – 850.
82. For a review on metallocene catalysis see: M. Hackmann, B. Rieger, *CatTech* **1** (1997) 79 – 92.
83. G. J. P. Britovsek, V. C. Gibson, D. F. Wass, *Angew. Chem.* **111** (1999) 448 – 468; *Angew. Chem. Int. Ed.* **38** (1999) 428 – 447.
84. S. D. Ittel, L. K. Johnson, M. Brookhart, *Chem. Rev.* **100** (2000) 1169 – 1203.
85. M. C. Baird, *Chem. Rev.* **100** (2000) 1471 – 1478.
86. L. S. Boffa, B. M. Novak, *Chem. Rev.* **100** (2000) 1479 – 1493.
87. T. R. Boussie, C. Coutard, H. W. Turner, V. Murphy, T. S. Powers, *Angew. Chem.* **100** (1998) 3472 – 3475; *Angew. Chem. Int. Ed.* **37** (1998) 3272 – 3275.
88. T. R. Boussie, V. Murphy, K. A. Hall, C. Coutard, C. Dales, M. Petro, E. Carlson, H. W. Turner, T. S. Powers, *Tetrahedron* **55** (1999) 11699 – 11710.
89. Symyx Technologies, US 6 306 658, 2000 (H. W. Turner, C. G. Dales, L. Van Erden, J. A. M. van Beek, D. A. Hadjuk, R. B. Nielson, P. Mansky, L. Matseiv).
90. Symyx Technologies, US 6 175 409, 1999 (R. B. Nielson, S. C. Kuebler, J. Bennett, A. Safir, M. Petro).
91. Symyx Technologies, US 6 030 917, 2000 and US 6 248 540, 2001 (W. H. Weinberg, E. W. McFarland, I. Goldwasser, T. Boussie, H. W. Turner, J. A. M. van Beek, V. Murphy, T. S. Powers).
92. M. Stork, A. Herrmann, T. Nemnich, M. Klapper, K. Müllen, *Angew. Chem.* **112** (2000) 4544 – 4547; *Angew. Chem. Int. Ed.* **39** (2000) 4367 – 4369.
93. J. Tian, G. W. Coates, *Angew. Chem.* **112** (2000) 3772 – 3773; *Angew. Chem. Int. Ed.* **39** (2000) 3626 – 3629.
94. C. Hinderling, P. Chen, *Angew. Chem.* **111** (1999) 2393 – 2396; *Angew. Chem. Int. Ed.* **38** (1999) 2253 – 2256.

95. For an overview on the chemistry of polyoxometalates see: C. L. Hill (ed.), *Chem. Rev.* **98** (1998) 1 – 390.
96. C. L. Hill, R. D. Gall, *J. Mol. Cat. A: Chem.* **114** (1996) 103 – 111.
97. E. Boring, Y. V. Geletti, C. L. Hill, *J. Am. Chem. Soc.* **123** (2001) 1625 – 1635.
98. P. Desrosiers, A. Guram, A. Hagemeyer, B. Jandeleit, D. M. Poojary, H. Turner, H. Weinberg, *Catal. Today* **67** (2001) 397 – 402.
99. For a comprehensive overview on heterogeneous catalysis see: G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **vol. 1 – 5,** VCH, Weinheim, 1997.
100. US 5 985 356, 1999 (P. G. Schultz, X.-D. Xiang, I. Goldwasser).
101. A. Holzwarth, P. Denton, H. Zanthoff, C. Mirodatos, *Catal. Today* **67** (2001) 309 – 318.
102. R. Schlögl, *Angew. Chem.* **110** (1998) 2467 – 2470; *Angew. Chem. Int. Ed.* **37** (1998) 2333 – 2336.
103. G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Preparation of solid catalysts,* Wiley-VCH, Weinheim, 1999.
104. J. M. Newsam, F. Schüth, *Combinatorial Chemistry Biotechnol. Bioeng.* **61** (1999) 203 – 216.
105. US 5 449 754, 1994 (G. M. Nashioka).
106. W. H. Weinberg, B. Jandeleit, K. Self, H. W. Turner, *Curr. Opin. Solid State Mater. Sci.* **3** (1998) 104 – 110.
107. B. E. Baker, N. J. Kline, P. J. Treado, M. J. Natan, *J. Am. Chem. Soc.* **118** (1996) 8721 – 8722.
108. E. Reddington, A. Sapienza, B. Gurau, R. Viswanathan, S. Sarangapani, E. S. Smotkin, T. E. Mallouk, *Science* **280** (1998) 1735 – 1737.
109. Symyx Technologies, WO 96/11878, 1996 (P. G. Schultz, X.-D. Xiang, I. Goldwasser).
110. E. Danielson, J. H. Golden, E. W. McFarland, C. M. Reaves, W. H. Weinberg, X. D. Wu, *Nature* **389** (1997) 944 – 948.
111. E. Danielson, M. Devenney, D. M. Giaquinta, J. H. Golden, R. C. Haushalter, E. W. McFarland, D. M. Poojary, C. M. Reaves, W. H. Weinberg, X. D. Wu, *Science* **279** (1998) 837 – 839.
112. S. M. Senkan, S. Öztürk, *Angew. Chem.* **111** (1999) 867 – 871; *Angew. Chem. Int. Ed.* **38** (1999) 791 – 795.
113. C. Hoffmann, A. Wolf, F. Schüth, *Angew. Chem.* **111** (1999) 2971 – 2975; *Angew. Chem. Int. Ed.* **38** (1999) 2800 – 2803.
114. K. E. Simons, 5th Heterogeneous Catalysis and Fine Chemicals, Lyon, 1999.
115. F. C. Moates, M. Somani, M. Annamalai, J. T. Richardson, D. Luss, R. C. Willson, *Ind. Eng. Chem. Res.* **34** (1996) 4801 – 4803.
116. A. Holzwarth, P. W. Schmidt, W. E. Maier, *Angew. Chem.* **110** (1998) 2788 – 2792; *Angew. Chem. Int. Ed.* **37** (1998) 2644 – 2647.
117. D. E. Akporiaye, I. M. Dahl, A. Karlsson, R. Wendelbo, *Angew. Chem.* **110** (1998) 629 – 631; *Angew. Chem. Int. Ed.* **37** (1998) 609 – 611.
118. J. Klein, C. W. Lehmann, H. W. Schmidt, W. F. Maier, *Angew. Chem.* **110** (1998) 3557 – 3561; *Angew. Chem. Int. Ed.* **37** (1998) 3369 – 3372.
119. P. Cong, R. D. Doolen, Q. Fan, D. M. Giaquinta, S. Guan, E. W. McFarland, D. M. Poojary, K. Self, H. W. Turner, W. H. Weinberg, *Angew. Chem.* **111** (1999) 507 – 512; *Angew. Chem. Int. Ed.* **38** (1999) 484 – 488.
120. M. Orschel, J. Klein, H. W. Schmidt, W. F. Maier, *Angew. Chem.* **111** (1999) 2961 – 2965; *Angew. Chem. Int. Ed.* **38** (1999) 2791 – 2803.
121. O. Buyevskaya, D. Wolf, M. Baerns, *Catal. Today* **62** (2000) 91 – 99.
122. U. Rodemerck, P. Ignaszewski, M. Lucas, P. Claus, M. Baerns, *Chem. Ing. Tech.* **71** (1999) 872.
123. P. Cong, A. Dehestani, R. Doolen, D. M. Giaquinta, S. Guan, V. Markov, D. Poojary, K. Self, H. W. Turner, W. H. Weinberg, *Proc. Natl. Acad. Sci. USA* **96** (1999) 11077.
124. Y. Liu, P. Cong, R. D. Doolen, H. W. Turner, W. H. Weinberg, *Catal. Today* **61** (2000) 87 – 92.
125. S. Senkan, K. Krantz, S. Öztürk, V. Zengin, I. Onal, *Angew. Chem.* **111** (1999) 2965 – 2971; *Angew. Chem. Int. Ed.* **38** (1999) 2794 – 2799.
126. U. Rodemerck, D. Wolf, O. V. Buyevskaya, P. Claus, S. Senkan, M. Baerns, *Chem. Eng. J.* **82** (2001) 3.
127. K. Krantz, S. Öztürk, S. Senkan, *Catal. Today* **62** (2000) 281 – 289.
128. H. Su, Y. Hou, R. S. Houk, G. L. Schrader, E. S. Yeung, *Anal. Chem.* **73** (2001) 4434 – 4440.
129. P. Claus, D. Hönicke, T. Zech, *Catal. Today* **67** (2001) 319 – 339.
130. G. Chen, D. A. Delafuente, S. Sarangapani, T. E. Mallouk, *Catal. Today* **67** (2001) 341 – 355.
131. O. V. Buyevskaya, A. Brückner, E. V. Kondratenko, D. Wolf, M. Baerns, *Catal. Today* **67** (2001) 369 – 378.

132. P. P. Pescarmona, J. C. van der Waal, I. E. Maxwell, T. Maschmeyer, *Angew. Chem.* **113** (2001) 762 – 765; *Angew. Chem. Int. Ed.* **40** (2001) 740 – 743.
133. C. Lettmann, H. Hinrichs, W. F. Maier, *Angew. Chem.* **113** ((2001) 3258 – 3262; *Angew. Chem. Int. Ed.* **40** (2001) 3160 – 3164.
134. E. M. Thorsteinson, T. P. Wilson, F. G. Young, P. H. Kasai, *J. Catal.* **52** (1978) 116.
135. US 4 250 346, 1981 (F. G. Young, E. M. Thorsteinson).
136. Y. Liu, P. Cong, R. Doolen, S. Guan, D. M. Poojary, H. W. Turner, W. H. Weinberg, 4th European Congress on Catalysis, Rimini, Italy, 1999, p. 117.
137. Y. Liu, P. Cong, R. Doolen, S. Guan, D. M. Poojary, H. W. Turner, W. H. Weinberg, 4th European Congress on Catalysis, Rimini, Italy, 1999, p. 41.
138. S. Senkan, *Nature* **394** (1998) 350 – 352.
139. D. Wolf, O. V. Buyevskaya, M. Baerns, *Appl. Catal. A* **200** (2000) 63 – 77.
140. B. Gurau, R. Viswanathan, R. Liu, T. J. Lafrenz, K. L. Ley, E. S. Smotkin, E. Reddington, A. Sapienza, B. C. Chan, T. E. Mallouk, S. Sarangapani, *J. Phys. Chem. B* **102** (1998) 9997 – 10003.
141. L. Liu, R. Viswanathan, R. Liu, E. S. Smotkin, *Electrochem. Solid State Lett.* **1** (1998) 123 – 125.
142. E. Reddington, J.-S. Yu, B. C. Chan, A. Sapienza, G. Chen, T. E. Mallouk, B. Gurau, R. Viswanathan, R. Liu, E. S. Smotkin, S. Sarangapani, *Comb. Chem.* 2000, 401 – 420.
143. P. Strasser, S. Gorer, M. Devenney, Proceedings of the International Symposium on Fuel Cells for Vehicles, 41st Battery Symposium, The Electrochemical Society of Japan, Nagoya 2000, p. 153.
144. M. G. Sullivan, H. Utomo, P. J. Fagan, M. D. Ward, *Anal. Chem.* **71** (1999) 4369 – 4375.
145. T. Siu, W. Li, A. K. Yudin, *J. Comb. Chem.* **2** (2000) 545 – 549.
146. T. Siu, S. Yekta, A. K. Yudin, *J. Am. Chem. Soc.* **122** (2000) 11787 – 11790.
147. A. K. Yudin, T. Siu, *Curr. Opin. Chem. Biol.* **5** (2001) 267 – 272.
148. K. Jambunathan, B. C. Shah, J. L. Hudson, A. C. Hillier, *J. Electroanal. Chem.* **500** (2001) 279 – 289.
149. B. C. Shah, A. C. Hillier, *J. Electrochem. Soc.* **147** (2000) 3043 – 3048.
150. M. T. Reetz, K. M. Kühling, S. Wilensek, H. Husmann, U. W. Hausig, M. Hermes, *Catal. Today* **67** (2001) 389 – 396.
151. L. Ma, X. Gong, E. S. Yeung, *Anal. Chem.* **72** (2000) 3383 – 3387.
152. Y. Zhang, X. Gong, H. Zhang, R. Larock, E. S. Yeung, *J. Comb. Chem.* **2** (2000) 450 – 452.
153. M. T. Reetz, K. M. Kühling, A. Deege, H. Hinrichs, D. Belder, *Angew. Chem.* **112** (2000) 4049 – 4052; *Angew. Chem. Int. Ed.* **39** (2000) 3891 – 3893.
154. S. A. Holmes, J. A. Saeedi, V. V. Buliants, P. Boolchand, D. Georgiev, U. Hacker, E. Sabkow, *Catal Today* **67** (2001) 403 – 409.
155. E. D. Isaacs, M. Marcus, G. Aeppli, X.-D. Xiang, X.-D. Sun, P. G. Schultz, M. Kao, G. S. Cargill III, R. Haushalter, *Appl. Phys. Lett.* **73** (1998) 1820 – 1822.
156. C. E. Kibbey in A. W. Czarnick, S. H. DeWitt (eds.): *A Practical Guide to Combinatorial Chemistry,* American Chemical Society, Washington D.C. 1997, pp. 123 – 151.
157. K. Omote, T. Kikuchi, J. Harada, M. Kawasaki, A. Ohtomo, M. Ohtani, T. Ohnishi, D. Komiyama, H. Koinuma, *Proc. SPIE-Int. Soc. Opt. Eng.* **3941** (2000) 84 – 91.
158. T. Bein, *Angew. Chem.* **111** (1999) 335 – 338; *Angew. Chem. Int. Ed.* **38** (1999) 323 – 326.
159. R. H. Crabtree, *Chem. Commun.* 1999, 1611 – 1616.
160. P. C. Pawlicki, R. A. Schmitz, *Chem. Eng. Prog.* **2** (1987) 40 – 45.
161. G. Georgiades, V. A. Self, P. A. Sermon, *Angew. Chem.* **99** (1987) 1050 – 1052; *Angew. Chem. Int. Ed.* **26** (1987) 1042 – 1044.
162. M. T. Reetz, M. H. Becker, M. Liebl, A. Fürstner, *Angew. Chem.* **112** (2000) 1294 – 1301; *Angew. Chem. Int. Ed.* **39** (2000) 1236 – 1239.
163. D. E. Bergbreiter, *Chemtracts: Org. Chem.* **10** (1997) 683 – 686.
164. S. Klein, S. Thorimbert, W. F. Meier, *J. Catal.* **163** (1996) 476 – 488.
165. S. Bukeikhanova, H. Orzsek, U. Kolb, K. Kühlein, W. F. Meier, *Catal. Lett.* **50** (1998) 93 – 105.
166. S. Storck, W. F. Meier, I. M. Miranda Salvado, J. M. F. Ferreira, D. Guhl, W. Souverijns, J. A. Martens, *J. Catal.* **172** (1997) 414 – 426.
167. W. J. Haap, T. B. Walk, G. Jung, *Angew. Chem.* **110** (1998) 3506 – 3509; *Angew. Chem. Int. Ed.* **37** (1998) 3311 – 3314.
168. C. M. Snively, G. Oskarsdottir, J. Lauterbach, *Angew. Chem.* **113** (2001) 3117 – 3120; *Angew. Chem. Int. Ed.* **40** (2001) 3028 – 3030.
169. C. M. Snively, S. Katzenberger, G. Oskarsdottir, J. Lauterbach, *Optics Letters* **24** (1999) 1841 – 1843.

170. C. M. Snively, G. Oskarsdottir, J. Lauterbach, *Catal. Today* **67** (2001) 357 – 368.
171. A. C. Cooper, L. H. McAlexander, D.-H. Lee, M. T. Torres, R. H. Crabtree, *J. Am. Chem. Soc.* **120** (1998) 9971 – 9972.
172. R. H. Crabtree, J. A. Loch, K. Gruet, D.-H. Lee, C. Borgmann, *J. Organomet. Chem.* **600** (2000) 7 – 11.
173. K. H. Shaughnessy, P. Kim, J. F. Hartwig, *J. Am. Chem. Soc.* **121** (1999) 2123 – 2132.
174. J. P. Stambuli, S. R. Stauffer, K. H. Shaughnessy, J. F. Hartwig, *J. Am. Chem. Soc.* **123** (2001) 2677 – 2678.
175. G. T. Copeland, S. J. Miller, *J. Am. Chem. Soc.* **121** (1999) 4306 – 4307.
176. A. H. Hoveyda, *Chemistry and Biology* **6** (1999) R305 – R308.
177. R. F. Harris, A. J. Nation, G. T. Copeland, S. J. Miller, *J. Am. Chem. Soc.* **122** (2000) 11270 – 11271.
178. O. Lavastre, J. P. Morken, *Angew. Chem.* **111** (1999) 3357 – 3359; *Angew. Chem. Int. Ed.* **38** (1999) 3163 – 3165.
179. A. Berkessel, R. Riedl, *J. Comb. Chem.* **2** (2000) 215 – 219.
180. K. Ding, A. Ishii, K. Mikami, *Angew. Chem.* **111** (1999) 519 – 522; *Angew. Chem. Int. Ed.* **38** (1999) 497 – 501.
181. K. Mikami, R. Angelaud, K. Ding, A. Ishii, A. Tanaka, N. Sawada, K. Kudo, M. Senda, *Chem. Eur. J.* **7** (2001) 730 – 737.
182. R. F. Service, *Science* **277** (1997) 474 – 475.
183. Symyx Technologies, WO 98/15969, 1997 (W. H. Weinberg, E. W. McFarland, P. Cong, S. Guan).
184. J. Gao, J. Wu, G. Siuzdak, M. G. Finn, *Angew. Chem.* **111** (1999) 1868 – 1871; *Angew. Chem. Int. Ed.* **38** (1999) 1755 – 1757.
185. M. T. Reetz, M. H. Becker, H.-W. Klein, D. Stöckigt, *Angew. Chem.* **111** (1999) 1872 – 1875; *Angew. Chem. Int. Ed.* **38** (1999) 1758 – 1761.
186. Y. Matsumoto, M. Murakami, Z. Jin, A. Nakayama, T. Yamaguchi, T. Ohmori, E. Suzuki, S. Nomura, M. Kawasaki, H. Koinuma, *Proc. SPIE-Int. Soc. Opt. Eng.* **3941** (2000) 19 – 27.
187. A. R. Connolly, J. D. Sutherland, *Angew. Chem.* **112** (2000) 4438 – 4441; *Angew. Chem. Int. Ed.* **39** (2000) 4268 – 4271.
188. Y. Yamada, A. Ueda, Z. Zhao, T. Maekawa, K. Suzuki, T. Takada, T. Kobayashi, *Catal. Today* **67** (2001) 379 – 387.
189. X.-D. Xiang, *Mater. Sci. Eng. B* **56** (1998) 246 – 250.
190. X.-D. Xiang, P. G. Schultz, *Physika C* 1997, 282 – 287, 428 – 430.
191. H. Koinuma, *Solid State Ionics* **108** (1998) 1 – 7.
192. X.-D. Sun, K.-A. Wang, Y. Yoo, W. G. Wallace-Freedman, C. Gao, X.-D. Xiang, P. G. Schultz, *Adv. Mater.* **9** (1997) 1046 – 1049.
193. B. E. Baker, N. J. Kline, P. J. Treado, M. J. Natan, *J. Am. Chem. Soc.* **118** (1998) 8721 – 8722.
194. T. Vossmeyer, S. Jia, E. DeIonno, M. R. Diehl, S.-H. Kin, X. Peng, A. P. Alivisatos, J. R. Heath, *J. Appl. Phys.* **84** (1998) 3664 – 3670.
195. T. X. Sun, *Biotechnol. Bioeng.* **61** (1999) 193 – 201.
196. R. B. van Dover, L. F. Schneemeyer, R. M. Fleming, H. A. Huggins, *Biotechnol. Bioeng.* **61** (1999) 217 – 225.
197. C. Schmitz, P. Poesch, M. Thelakkat, H.-W. Schmidt, *Proc. SPIE-Int. Soc. Opt. Eng.* **3797** (1999) 423 – 431.
198. H.-W. Schmidt, C. Schmitz, P. Poesch, M. Thelakkat, *Proc. SPIE-Int. Soc. Opt. Eng.* **3797** (1999) 58 – 65.
199. C. Schmitz, M. Thelakkat, H.-W. Schmidt, *Adv. Mater.* **11** (1999) 821 – 826.
200. C. Schmitz, P. Posch, M. Thelakkat, H.-W. Schmidt, *Phys. Chem. Chem. Phys.* **1** (1999) 1777 – 1781.
201. C. Schmitz, P. Posch, M. Thelakkat, H.-W. Schmidt, *Macromol. Symp.* **154** (2000) 209 – 221.
202. E. W. McFarland, W. H. Weinberg, *Trends Biotechnol.* **17** (1999) 107 – 115.
203. E. W. McFarland, W. H. Weinberg, *Mater. Technol.* **3** (1998) 153 – 158.
204. S. Pessaud, F. Gervais, C. Champeaux, P. Marchet, A. Catherinot, M. Licheron, J. L. Longuet, F. Ravel, *Mat. Sci. Eng. B* **B60** (1999) 205 – 211.
205. J. Letteri, Y. Jia, S. J. Fulk, D. G. Schlom, M. E. Hawley, G. W. Brown, *Thin Solid Films* **379** (2000) 64 – 71.
206. C.-M. Chen, H. C. Pan, D. Z. Zhu, J. Hu, M. Q. Li, *Nucl. Instrum. Methods Phys. Res., Sect. B* **150** (1999) 81 – 88.
207. X. Q. Liu, Z. F. Li, W. Lu, S. C. Chen, C. M. Chen, D. Z. Zhu, J. Hu, M. Q. Li, *Appl. Phys. Lett.* **75** (1999) 2611 – 2613.
208. C. Chen, H. Pan, D. Zhu, J. Hu, M. Li, *Gaojishu Tongxun* **9** (1999) 41 – 44.
209. C. Chen, H. Pan, D. Zhu, J. Hu, M. Li, *Mater. Sci. Eng. B* **B72** (2000) 113 – 116.

210. H. Koinuma, T. Koida, T. Ohnishi, D. Komiyama, M. Lippmaa, M. Kawasaki, *Appl. Phys. A.: Mater, Sci. Process., Suppl.* **69** (1999) S29 – S31.
211. A. Ohtomo, T. Makino, K. Tamura, Y. Matsumoto, Y. Segawa, Z. Tang, G. K. L. Wong, H. Kionuma, M. Kawasaki, *Proc. SPIE-Int. Soc. Opt. Eng.* **3941** (2000) 70 – 81.
212. Y. Matsumoto, M. Murakami, Z. Jin, A. Nakayama, T. Yamaguchi, T. Ohmori, E. Suzuki, S. Nomura, M. Kawasaki, H. Koinuma, *Proc. SPIE-Int. Soc. Opt. Eng.* **3941** (2000) 19 – 27.
213. Z. Jin, M. Murakami, T. Fukumura, Y. Matsumoto, A. Ohtomo, M. Kawasaki, H. Koinuma, *J. Cryst. Growth* **214/215** (2000) 55 – 58.
214. H. Aiyer, D. Nishioka, R. Maruyama, H. Shinno, N. Matsuki, K. Miyazaki, H. Fujioka, H. Koinuma, *Jpn. J. Appl. Phys.* **40** (2001) L81 – L83.
215. J. Wang, Y. Yoo, C. Gao, I. Takeuchi, X.-D. Sun, H. Chang, X.-D. Xiang, P. G. Schultz, *Science* **279** (1998) 1712 – 1714.
216. Symyx Technologies, WO-A 98/14641, 1998 (E. W. McFarland, E. Danielson, M. Devenney, C. Warren).
217. X.-D. Xiang, *Chem. Ind.* **19** (1998) 800 – 802.
218. J. Wang, Y. Yoo, C. Gao, I. Takeuchi, X.-D. Sun, H. Chang, X.-D. Xiang, P. G. Schultz, *Science* **279** (1998) 1712 – 1714.
219. For a review, see: M. Noh, C. D. Johnson, M. D. Hornbostel, J. Thiel, D. C. Johnson, *Chem. Mater.* **8** (1996) 1625 – 1635.
220. C. D. Johnson, *Chem. Mater.* **10** (1998) 1096 – 1101.
221. J. L. Wu, M. Devenney, E. Danielson, D. Poojary, W. H. Weinberg, *Mater. Res. Soc. Symp. Proc.* **560** (1999) 65 – 70.
222. J. T. Rantala, T. Kololuma, L. Kivimaki, *SPIE-Int. Soc. Opt. Eng.* **3941** (2000) 11 – 18.
223. J. Wang, K. E. Gonsalves, *J. Comb. Chem.* **1** (1999) 216 – 222.
224. X.-D. Sun, K.-A. Wang, Y. Yoo, W. G. Wallace-Freedman, C. Gao, X.-D. Xiang, P. G. Schultz, *Adv. Mater.* **9** (1997) 1046 – 1049.
225. D. E. Akporiaye, I. M. Dahl, A. Karlsson, R. Wendlbo, *Angew. Chem.* **110** (1998) 629 – 631; *Angew. Chem. Int. Ed.* **37** (1998) 609 – 611.
226. J. Klein, C. W. Lehmann, H.-W. Schmidt, W. F. Maier, *Angew. Chem.* **110** (1998) 3557 – 3561; *Angew. Chem. Int. Ed.* **37** (1998) 3369 – 3372.
227. K. Choi, D. Gardner, N. Hildbrandt, T. Bein, *Angew. Chem.* **111** (1999) 3070 – 3073; *Angew. Chem. Int. Ed.* **38** (1999) 2891 – 2894.
228. B. E. Baker, N. J. Kline, P. J. Treado, M. J. Natan, *J. Am. Chem. Soc.* **118** (1996) 8721 – 8722.
229. K. Kneipp, H. Kneipp, R. Manoharan, E. B. Hanlon, I. Itzkan, R. R. Dasari, M. S. Field, *Appl. Spectrosc.* **52** (1998) 1493 – 1497.
230. J. M. Whitling, G. Spreitzer, D. W. Wright, *Adv. Mater.* **12** (2000) 1377 – 1380.
231. G. Spreitzer, J. M. Whitling, D. W. Wright, *Mater. Res. Soc. Symp. Proc.* **599** (2000) 201 – 206.
232. G. Spreitzer, J. M. Whitling, J. D. Madura, D. W. Wright, *Chem. Commun.* 2000 209 – 210.
233. M. Cohen-Adad, M. Gharbi, C. Goutaudier, R. Cohen-Adad, *J. Alloys Compd.* **289** (1999) 185 – 195.
234. S. Inoue, *Kagaku Kogyo* **51** (2000) 128 – 133.
235. E. D. Isaacs, M. Marcus, G. Aeppli, X.-D. Xiang, X.-D. Sun, P. G. Schultz, M. Kao, G. S. Cargill III, R. Haushalter, *Appl. Phys. Lett.* **73** (1998) 1820 – 1822.
236. K. Omote, T. Kikuchi, J. Harada, M. Kawasaki, A. Ohtomo, M. Ohtani, T. Ohnishi, D. Komiyama, H. Koinuma, *Proc. SPIE-Int. Soc. Opt. Eng.* **3941** (2000) 84 – 91.
237. T. Makino, G. Isoya, Y. Segawa, C. Chia, T. Yasuda, M. Kawasaki, A. Ohtomo, K. Tamura, Y. Matsumoto, H. Koinuma, *Proc. SPIE-Int. Soc. Opt. Eng.* **3941** (2000) 28 – 35.
238. J. G. Bednorz, M. Takashige, K. A. Müller, *Europhys. Lett.* **3** (1987) 379 – 385.
239. X.-D. Xiang, P. G. Schultz, *Physica C* **282 – 287** (1997) 428 – 430.
240. X.-D. Xiang, *Biotechnol. Bioeng.* **61** (1999) 227 – 241.
241. X.-D. Xiang, *Annu. Rev. Mater. Sci.* **29** (1999) 149 – 171.
242. C. Gao, T. Wei, F. Duewer, X.-D. Xiang, *Appl. Phys. Lett.* **71** (1997) 1817 – 1821.
243. A. W. Webb, E. F. Skelton, S. B. Quadri, V. Browning, E. R. Carpenter, *Physical Review B* **45** (1992) 2480 – 2483.
244. A. W. Webb, E. F. Skelton, S. B. Quadri, E. R. Carpenter, *J. Solid State Chem.* **102** (1993) 519 – 533.
245. I. Bozovic, V. Matijasevic, *Mat. Sci. Forum* **352** (2000) 1 – 8.
246. H. Kionuma, Y. Matsumoto, M. Murakami, T. Shono, T. Hasegawa, T. Fukumura, M. Kawasaki, P. Ahmet, T. Chikyow, S.-Y. Kioshihara, *Science* **201** (2001) 854 – 856.

247. Y. Matsumoto, M. Murakami, Z. Jin, A. Nakayama, T. Yamaguchi, T. Ohmori, E. Suzuki, S. Nomura, M. Kawasaki, H. Koinuma, *Proc. SPIE-Int. Soc. Opt. Eng.* **3941** (2000) 19 – 27.
248. G. Briceno, H. Chang, X.-D. Sun, P. G. Schultz, X.-D. Xiang, *Science* **270** (1995) 273 – 275.
249. Symyx Technologies, US 5 776 359, 1998 (P. G. Schultz, X.-D. Xiang, I. Goldwasser, The Regents of the University of California).
250. K. Seeger: *Semiconductor Physics,* Springer, Berlin, 1991.
251. R. B. van Dover, L. F. Schneemeyer, R. M. Fleming, H. A. Huggins, *Biotechnol. Bioeng.* **61** (1999) 217 – 225.
252. J. J. Hanak, *Proc. 1er Colloq. Int. Pulvérisation Cathodiques Ses Appl. Suppl.* **165** (1973) 177 – 197.
253. R. B. van Dover, L. F. Schneemeyer, R. M. Fleming, *Nature* **392** (1998) 162 – 164.
254. H. Chang, C. Gao, I. Takeuchi, Y. Yoo, J. Wang, P. G. Schultz, X.-D. Xiang, R. P. Sharma, M. Downes, T. Venkatesan, *Appl. Phys. Lett.* **72** (1998) 2185 – 2187.
255. J. Li, F. Duewer, C. Gao, H. Chang, X.-D. Xiang, Y. Lu, *Appl. Phys. Lett.* **76** (2000) 769 – 771.
256. X.-D. Xiang, *Mater. Sci. Eng. B* **56** (1998) 246 – 250.
257. E. W. McFarland, W. H. Weinberg, *Mater. Technol.* **13** (1998) 107 – 120.
258. Y. Lu, T. Wei, F. Duewer, Y. Lu, N.-B. Ming, P. G. Schultz, X.-D. Xiang, *Science* **276** (1997) 2004 – 2006.
259. T. Wei, W. G. Wallace-Freedman, P. G. Schultz, X.-D. Xiang, *Appl. Phys. Lett.* **68** (1996) 3506 – 3508.
260. C. Gao, T. Wei, F. Duewer, Y. Lu, X.-D. Xiang, *Appl. Phys.* **71** (1997) 1872 – 1874.
261. I. Takeuchi, F. Duewer, X.-D. Xiang, *Appl. Phys.* **71** (1997) 2026 – 2028.
262. I. Takeushi, H. Chang, C. Gao, P. G. Schultz, X.-D. Xiang, R. P. Sharma, M. J. Downes, T. Venkatsan, *Appl. Phys. Lett.* **73** (1998) 894 – 896.
263. T. Jüstel, H. Nikol, C. Ronda, *Angew. Chem.* **110** (1998) 3250 – 3271; *Angew. Chem. Int. Ed.* **37** (1998) 3084 – 3103.
264. X.-D. Sun, C. Gao, J. Wang, X.-D. Xiang, *Appl. Phys. Lett.* **70** (1997) 3353 – 3355.
265. X.-D. Sun, K.-A. Wang, Y. Yoo, W. G. Wallace-Freedman, C. Gao, X.-D. Xiang, P. G. Schultz, *Adv. Mater.* **9** (1997) 1046 – 1049.
266. X.-D. Sun, X.-D. Xiang, *Appl. Phys. Lett.* **72** (1998) 525 – 527.
267. C. Schmitz, H.-W. Schmidt, M. Thelakkat, *Chem. Mater.* **12** (2000) 3012 – 3019.
268. V. N. Nesterov, T. V. Timofeeva, O. Y. Borbulevych, M. Y. Antipin, R. D. Clark, *Acta Cryst.* **C56** (2000) 971 – 975.
269. K. Sada, K. Yoshikawa, M. Miyata, *Chem. Commun.* 1998 1763 – 1764.
270. Q. Huo, G. Sui, P. Kele, R. M. Leblanc, *Angew. Chem.* **112** (2000) 1934 – 1936; *Angew. Chem. Int. Ed.* **39** (2000) 1854 – 1857.
271. For reviews about new chiral selectors see: A. F. Svec, D. Wulff, J. M. J. Fréchet in G. Subramanian (ed.): *Chiral Separation Techniques,* Wiley-VCH, Weinheim, 2001, pp. 55 – 93.
272. P. Murer, K. Lewandoski, F. Svec, J. M. J. Fréchet, *Anal. Chem.* **71** (1999) 1278 – 1284.
273. K. Lewandowski, P. Murer, F. Svec, J. M. J. Fréchet, *J. Comb. Chem.* **1** (1999) 105 – 112.
274. Y. Wu, Y. Wang, A. Yang, T. Li, *Anal. Chem.* **71** (1999) 1688 – 1691.
275. L. H. Bluhm, Y. Wang, T. Li, *Anal. Chem.* **72** (2000) 5201 – 5205.
276. C. J. Welch, *J. Chromatogr. A* **666** (1994) 3 – 26.
277. C. J. Welch, G. Bhat, M. N. Protopopova, *J. Combi. Chem.* **1** (1999) 364 – 367.
278. Y. Wang, T. Li, *Anal. Chem.* **71** (1999) 4178 – 4182.
279. Y. Wang, L. H. Bluhm, T. Li, *Anal. Chem.* **72** (2000) 5459 – 5465.
280. C. J. Welch, S. D. Pollard, D. J. Mathre, P. J. Reider, *Org. Lett.* **3** (2000) 95 – 98.
281. M. D. Weingarten, K. Sekanina, W. C. Still, *J. Am. Chem. Soc.* **120** (1998) 9112 – 9113.
282. O. Ramström, L. Ye, M. Krook, K. Mosbach, *Anal. Comm.* **35** (1998) 9 – 11.
283. O. Ramström, L. Ye, M. Krook, K. Mosbach, *Chromatographia* **47** (1998) 465 – 469.
284. L. Sabourin, R. J. Ansell, K. Mosbach, I. A. Nicholls, *Anal. Comm.* **35** (1998) 285 – 287.
285. For a review on molecular imprinting see: G. Wulff, *Angew. Chem.* **107** (1995) 1958 – 1979; *Angew. Chem. Int. Ed.* **34** (1995) 1812 – 1832.
286. T. Takeuchi, D. Fukuma, J. Matsui, *Anal. Chem.* **71** (1999) 285 – 290.
287. S. Brocchini, K. James, V. Tangpasuthadol, J. Kohn, *J. Am. Chem. Soc.* **119** (1997) 4553 – 4554.
288. T. A. Dickinson, D. R. Walt, J. White, J. S. Kauer, *Anal. Chem.* **69** (1997) 3413 – 3418.
289. D. J. Gravert, A. Datta, P. Wentworth Jr., K. D. Janda, *J. Am. Chem. Soc.* **120** (1998) 9481 – 9495.

290. A. W. Bosman, A. Heuman, G. Klärner, D. Benoit, J. M. J. Fréchet, C. J. Hawker, *J. Am. Chem. Soc.* **123** (2001) 6461 – 6462.
291. C. J. Hawker, D. Benoit, F. Rivera Jr., V. Chaplinski, A. Nilsen, R. Braslau, *Polym. Mater. Sci. Eng.* **80** (1999) 90 – 91.
292. R. B. Nielsen, A. L. Safir, M. Petro, T. S. Lee, P. Huefner, *Polym. Mater. Sci. Eng.* **80** (1999) 92.
293. G. Klärner, A. L. Safir, H.-T. Chang, M. Petro, R. B. Nielsen, *Polym. Prepr.* **401** (1999) 469.
294. T. Hashimoto, Y. Muramatsu, T. Hayakawa, H. Fukumoto, T. Yamamoto, H. Koinuma, *Proc. SPIE-Int. Soc. Opt. Eng.* **3941** (2000) 36 – 44, 92 – 100.

Dry Mortars

Roland Bayer, Wolff Cellulosics GmbH & Co KG, D-29699 Bomlitz, Germany

Hermann Lutz, Wacker Polymer Systems GmbH & Co KG, D-84489 Burghausen, Germany

1. Introduction 1
2. Development 2
2.1. Historical and Technical Development 2
2.2. Advantages of Dry-Mix Mortars . . 2
3. Composition 4
3.1. Binders 4
3.1.1. Mineral Binders 5
3.1.2. Organic Binders 5
3.2. Aggregates 6
3.3. Additives 6
3.3.1. Cellulose Ethers 7
3.3.2. Other Additives 8
4. Production 10
5. Testing 11
6. Applications 13
6.1. Brick-Laying Mortars and Related Adhesives 13
6.2. Renders and Plasters 13
6.3. Tile Adhesives 16
6.4. Tile Grouts 18
6.5. Exterior Thermal Insulation Composite Systems 19
6.6. Powder Paints 21
6.7. Cementitious Waterproofing Sealing Slurries 22
6.8. Self-Leveling Underlayments and Screeds 23
6.9. Patching and Repair Mortars 24
7. Market Aspects 25
8. References 25

1. Introduction

Mortars based on mineral binders like lime, cement, or gypsum have been used for more than 8000 years in the construction of buildings. These mortars have mainly been used for laying stones and bricks (masonry mortars) and for coating walls (rendering mortars). Until the 1950s cement-based mineral mortars were exclusively produced and applied by so-called job-site mixing technology. Job-site mixing means transportation of the individual raw materials to the job-site and their mixing on site in the appropriate ratio. Thus cement, the most common mineral binder, is mixed with fillers (sand) before water is added to create the wet mortar for application.

Similar to the way in which job-site-mixed concrete was substituted by the economically and ecologically more favorable ready-mix concrete, job-site mixing techology for masonry and rendering mortars was replaced by factory-mixed dry mortars, also called dry-mix mortars. Dry-mix mortars or dry mortars are produced in specially designed dry-mix mortar plants in which mineral binder(s) and aggregates (sand) are mixed together in the appropriate way. This factory-based process also allows different additives and admixes to be added to these dry mortars to improve significantly their technical performance. Based on this technology individual dry mortars for specific applications can be produced according to formulations developed and pretested in the laboratory.

The factory-mixed dry mortars are delivered to the construction site in bags or in special silos and need only be mixed with water prior to use. Together with the appropriate equipment for efficient transport, mixing with water, and machine application of the wet mortar, this dry-mix mortar technology led to a drastic improvement in productivity in the application of high-volume products like masonry and rendering mortars.

The possibility of adding specific dry additives or admixes in a well defined ratio to the dry mix during the production also led to the development of high-quality mineral mortars with well-defined and specific technical properties. These highly specialized mortars corresponding to the requirements of the modern building industry can not be produced with job-site

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.f16_f01

mortar technology. Consequently, high-quality, additive- and admix-modified mineral mortars are today widely used in the building industry and have largely substituted other building materials such as ready-to-use paste compounds and liquid admixes used in combination with mineral mortars.

2. Development

2.1. Historical and Technical Development

For thousands of years the architecture and construction of buildings were closely associated with the use of mineral mortars. Lime plasters have been known for more than 8000 years, and gypsum mortars were used by the Babylonians about 6000 years ago. Hydraulic-setting mortars based on pozzolans (powdered volcanic ash) have probably been known for over 3000 years and were used in large amounts by the ancient Phoenicians, Greeks, and Romans.

In antiquity and the Middle Ages, additives and admixes such as soaps, resins, proteins, and ash were already mixed on the job site with mineral binders and aggregates to improve the technical performance of the resulting mortars.

Although a first patent on the manufacture and application of dry-mix mortars was published as early as 1893 in Europe, mortars were applied up to the 1950s exclusively as job-site-mixed mortars. For these mortars, the mineral binders (mostly cement) and aggregates (mostly quartz sand) are transported separately to the job site and are then mixed together by hand in the appropriate ratio. After mixing with water, the wet mortar is ready for application.

During the 1950s and 1960s in Western Europe and in the USA, but especially in Germany, there was a fast-growing demand in the construction industry for new building materials and technologies. The reasons for this included shortage of skilled workmen, the need for shorter construction times together with cost reduction, increasing labor costs, the diversification of building materials suitable for specific applications, new materials, and an increased demand for better quality of construction.

Job-site mortar technology was and is not able to adequately meet all these requirements. As a practical consequence the development of the modern construction and building chemical industry in the countries of the Western world from the 1960s onwards was influenced mainly by three important trends, which can be seen nowadays worldwide:

- Replacement of the job-site-mixed mortars by premixed and prepacked dry-mix mortars.
- Mechanization of mortar application, including bulk transportation systems (e.g., silos), mechanical systems for automatic mixing of dry-mix mortar with water, and machine application (spraying) of wet mortar.
- Modification of mortars with polymer binders (redispersible powders) and special additives (e.g., cellulose ethers) and admixes to improve product quality and to meet the requirements of the modern building industry.

The introduction of dry-mix mortar techology and the use of silo transport and machine appliation of mortars made it possible that from 1960 to 1995 the volume of render and plaster mortar application in Germany increased by 600 %, while the number of employees in this sector decreased by 25 %, that means, productivity increased by 800 % [1].

2.2. Advantages of Dry-Mix Mortars

In job-site mixing technology, for each building application a specific ratio of cement and fillers is mixed appropriately before being gauged with water, and the wet mortar is applied. The quality of such a mortar depends on the quality of the raw materials, their correct mixing ratio, the homogeneity of the mixture, and the consistency of the fresh mortar. Under these conditions the quality of mortars produced by job-site mix technology can not be guaranteed. The main disadvantages of job-site-mixed mortars are that the whole process can not be automated, and the producer and applicator of a job-site-mixed mortar can not give a warranty to customers and end users due to a high risk of inconsistent quality of the prepared mortars. In addition, additives can either not be added or can only be added with a high risk of dosage and mixing errors and of obtaining inhomogeneous mixtures. The possibilities for producing specialized and individual

products with job-site technology are very limited. Last but not least, handling and the logistics of job-site-mixed mortars are more complicated, and their applications are very limited.

In contrast to construction-site mortars, the modern dry-mix mortars are produced in a dry-mix mortar plant by mixing together all necessary ingredients such as binders, aggregates, and chemical additives. In this way different kinds of dry-mix mortars with well-defined product characteristics that correspond to the requirements for specific applications can be produced. The use of premixed and prepacked dry-mix mortars not only increases significantly production performance and the productivity on construction sites, but also guarantees a high degree of application safety and reliability in avoiding on-site mixing errors. Prepacked dry-mix mortars produced in a dry-mix plant assure that binders, aggregates, and additives of high and constant quality are mixed exactly in the same ratio, thus ensuring a consistent high level of quality [2], [3].

The raw materials used for the production of prepacked dry-mix mortars can be classified as follows:

1) Mineral binders
   a) Portland cement (OPC)
   b) High-alumina cement (HAC)
   c) Special cements
   d) Hydrated lime
   e) Gypsum
   f) Anhydrite
2) Polymer binder (redispersible powder)
3) Aggregates, fillers
   a) Silica sands
   b) Limestone sand
   c) Dolomite sand
   d) Marble sand
   e) Lightweight fillers
   f) Special and functional fillers
4) Additives
   a) Cellulose ethers
   b) Pigments
   c) Defoaming agents
   d) Air-entraining agents
   e) Retarding agents
   f) Accelerators
   g) Thickening agents
   h) Hydrophobing agents
   i) Plasticizing agents
   j) Superplasticizer

The main applications of dry-mix mortars can be classified as in the following:

1) High-volume products (ca. 70 % of total dry-mix mortar volume produced)
   – Masonry mortars
   – Base renders (cement- and gypsum-based)
   – Bricklaying mortars
   – Bricklaying adhesives
   – Cement-based screeds
   – Gypsum-based screeds
   – Dry concrete
   – Shotcrete concrete
   – Mineral plasters
2) Specialized products (ca. 30 % of total dry-mix mortar volume produced)
   – Ceramic-tile adhesives
   – Building adhesives
   – Tile grouts
   – Grouts
   – Decorative mineral plasters
   – Powder paints
   – External thermal composite insulation systems
   – Trowelling compounds
   – Flooring compounds
   – Repair mortars

The two different approaches to produce fresh, wet mortars ready for application either as a job-site mixed mortar (process A) or as a factory-made prepacked dry-mix mortar (process B) have major consequences for handling and productivity, as exemplified for base-render application in Table 1.

The application of high-volume dry-mix mortar was strongly promoted by the development of bulk transportation containers (silo systems) and mechanical systems for gauging the mortars with water and pumping for mechanical application by spraying. Compared to the manual gauging of dry-mix mortars delivered in bags to the job site (process B), the use of automatic gauging and pumping devices for the mechanical application of the mortar leads to an additional improvement in productivity (process D).

Table 1. Comparison of production, transport, and application of job-site-mixed mortars and dry-mix mortars

Process∗	A	B	C	D
Transport of sand and cement separately to the job-site	+			
Manual mixing of the mineral binder and the aggregates on the job site	+			
Transport of the prepacked dry-mix mortar in bags to the job site		+	+	
Transport of the prepacked dry-mix mortar in silos to the job site				+
Manual mixing of the mortar with water	+	+		
Manual application of the wet mortar	+	+		
Machine mixing of the dry-mix mortar with water and machine spray application			+	+
Productivity (m^2 per man shift) for a render application	10	25	40	50 – 60

∗ A: completely manual method (job-site mixing technique); B, C, D: dry-mix mortar production in a dry-mix plant with different methods of transport and application.

The handling of dry-mix mortars in bags could be eliminated for many high-consumption applications by filling the dry-mix mortars produced in the dry-mix plant into containers 1 – 20 m^3 in volume which are transported to the site. With appropriate conveying systems the dry mortar is transferred directly from the silo into the attached mixing and pumping unit, where it is gauged automatically with water and pumped for spraying.

Combining the transportation of the dry-mix mortar in silos or containers with automatic gauging, pumping, and mechanical application of the mortar provides a further improvement in productivity (process D). In addition to the improved productivity, the automatic mechanical gauging and spraying of dry-mix mortars ensures consistency in handling and application of these products. Possible errors such as over- or underdosage of water or incorrect composition of the mortars are eliminated, which is especially important if less experienced or unskilled workers are working on the job site.

In Western Europe the consequences of this development were phenomenal. Since the 1960s a huge number of modern dry mortar plants have been established with millions of tonnes of capacity. In Germany, for example, nowadays approximately 100 dry mortar plants exist, producing 10×10^6 t/a of dry-mix mortars. There was a tremendous boom in dry-mix mortar technology after reunification of Germany after 1990, which now continues in the countries of Eastern Europe. The average growth rate for dry-mix mortar applications in Europe is approximately 12 % per annum, based on a production of about $35 – 40 \times 10^6$ t/a in 2000 [4].

3. Composition

Dry mortars are generally composed at least of three components: a binder, an aggregate, and additives.

Modern dry mortars are composed of many more components than those which were mixed directly on the site in the past. Today the simplest formulations are masonry mortars, brick-laying mortars, and low-quality tile adhesives, while highly sophisticated and high-performance dry mortars like self-leveling floor underlayments and decorative plasters may contain up to 20 ingredients. In this article the word aggregates is used for all types of mineral ingredients which do not have a binder function, even if they are added in small amounts like mineral ingredients with special functions, e.g., functional fibers or pigments. Other ingredients with small addition rates (redispersible powders) are treated as organic binders, although many others describe them as additives. The effect of using redispersible powders, especially in tile adhesives, shows clearly that they have a binder function. On the other hand, methyl cellulose, a multipurpose additive, is described in the chapter on additives although it also has a certain binder effect. In this case the binder effect does not play the major role.

3.1. Binders

The binder glues aggregates and other particles together and provides adhesion to the substrate. Due to their physical or chemical reaction binders play the major role for the final strength of the mortar. Binders can be classified as hydraulic and nonhydraulic binders. The former also set under water.

The setting reaction of the binders cement and hydrate lime occurs by chemical reaction. Cement reacts on contact with water during mixing while hydrate lime (without hydraulic portions) sets by reaction with atmospheric carbon dioxide. Gypsum and organic binders set physically: gypsum by recrystallization with water and formation of matted crystal needles, and organic binders by forming a homogeneous polymer film.

3.1.1. Mineral Binders

Cement (see also → Cement and Concrete). In dry mortars ordinary portland cement (OPC) is mainly used. The hydration reaction leads primarily to the formation of calcium silicate hydrates, which retain their strength and stability even under water (hydraulic binder). While the quality of an OPC like CEM I 32.5R (for nomenclature, see DIN 1164 [5]) is sufficient for bricklaying mortars, masonry cement, renders, and many plasters, in tile adhesives higher qualities like CEM I 42.5R or CEM I 52.5R are desired. Dry mortars that are both mortars and decorative finishes, such as decorative plasters and tile grouts, mostly contain white portland cement. Fast-setting high-alumina cement (e.g., Fondue Lafarge) consists mainly of calcium aluminates and is used for dry mortars that require fast-setting properties or high-temperature stability.

Gypsum [6] (see also → Calcium Sulfate). Calcium sulfate hemihydrate and anhydrite both set with water by forming calcium sulfate dihydrate.

Calcium sulfate hemihydrate exists in two crystalline forms (depending on the production process): the α-form with larger crystals, higher tensile and compressive strength, and lower water demand, and the more amorphous β-form with high porosity and lower tensile and compressive strength and an up to three times higher water demand.

Anhydrite exists in two application-relevant phases: the anhydrite II phase, which is of importance in anhydrite-based screeds and the anhydrite III phase, which is a part of the multiphase gypsum used for plasters. Gypsum plasters contain β-hemihydrate as well as anhydrite II and III. Gypsum-based joint filler contains β-semihydrate.

Hydrated Lime (see also → Lime and Limestone), [7]. Hydrated lime sets by reaction with carbon dioxide to form calcium carbonate and is therefore not a hydraulic binder. Hydraulic properties in some hydrated lime result from impurities or added materials with pozzolanic properties. For centuries hydrated lime was by far the most important binder in mortars. Today it is largely substituted by hydraulic binders, which set faster, but it is still in use, mainly because of its plastic properties. The workability of many dry mortars is improved by adding 5 – 30 wt % hydrated lime to the cement. Limes (including powdered hydrated lime) for building purposes are specified in DIN EN 459 [8].

3.1.2. Organic Binders

Improving the characteristics of a cementitious mortar with organic materials is well known. In antiquity, for example, proteins in the form of liquid milk or even blood were used. In most modern applications, mortars unmodified by organic polymer binders are no longer able to meet state-of-the art technical requirements. Even cementitious mortars that contain cellulose ethers as an additive to improve their water-retention capability and workability characteristics adhere poorly, or not at all, to many of the materials used in the modern construction industry (e.g., polystyrene, cement fiber and wood panel; non-absorbent substrates like old tiles and fully vitrified tiles). Furthermore, cementitious mortars are very hard, brittle, and inflexible materials, but for many applications flexible and deformable cementitious mortars must be used. Thus, for many applications in the modern construction industry the modification of cementitious mortars with polymers today is a must. In two-binder systems, the mineral binder cement and the polymer binder in the form of a redispersible powder complement each other ideally. Their combination results in outstanding synergistic properties and characteristics of the dry-mix mortar which cannot be produced by either of the two individual binders alone.

In the early 1930s water-based liquid dispersions [9] were added prior to or together with the gauging water if required. Mortars which are modified in this way are referred to as two-pack systems (powder-form mineral binder plus liquid polymer binder in a second pack). However, in practice many faults occurred with the use of the two-pack (or two-component) systems at the construction site. The main difficulty is the precise dosing of the polymer dispersion in its liquid form.

Dosing errors may occur due to insufficient knowledge, experience, and training of workers concerning the appropriate dosage for a specific application with precise requirements, or due to incorrect dosage, which may happen unintentionally by error or even intentionally to save on short-term costs. Incorrect dosage of the liquid polymer dispersion will change the characteristics and the technical performance of the mortar significantly, and this will lead to severe damages in various construction materials, e.g., through insufficient adhesive bond strength, flexibility and/or durability. Other reasons that argue against two-pack systems, apart from the difficult and risky handling, are the additional expenses and logistics difficulties (e.g., the need for additional containers and their subsequent disposal, storage and transport of liquid dispersions which could freeze or deteriorate by microbiological attack, more time consuming and ponderous handling on the job site with the two-pack system).

The invention of redispersible powders (trade name Vinnapas® Redispersible Powder) by Wacker Chemie in 1953 made possible the production of the first polymer-modified dry-mix mortars, known today as one-pack or one-component systems. Redispersible powders are polymer binding agents produced by spray-drying special water-based dispersions, mostly based on vinyl acetate – ethylene copolymers. These are often also referred to as redispersible powders [10], [11], because after mixing or redispersion with water, these powder polymer binders can be returned to their original water-based dispersion with all their typical characteristics and functions as polymer binders. The polymer film acting as a binder is formed after partial evaporation of the water by coalescence of the individual polymer particles. This polymer film acts as an organic binder, gluing together the filler particles, reinforcing the mortar structure and providing an excellent adhesion at the mortar – substrate interface. Polymer films in a cement mortar are shown in Figure 1.

The use of factory-made dry-mix mortars with precisely stipulated proportions for cement, aggregates, additives, and redispersible powders as organic binder results in a high-quality product with a high degree of application safety by avoiding possible errors during dosing and mixing at the construction site.

The modification of dry-mix mortars with redispersible polymer powders improves, depending on the dosage, the adhesive bond strength on all kind of substrates, the flexibility and deformability of the mortars, the flexural strength, and the abrasion resistance, the toughness, the cohesion and the density (impermeability) of the mortar as well as the water retentivity and the workability characteristics. In addition special redispersible powders with a hydrophobing effect can result in a strong water-repellent effect of the mortars.

3.2. Aggregates

The majority of aggregates are normal size fractions of quartz, limestone, or dolomite. To adjust the grain-size distribution (see Section 5) normally different granulometric fractions of aggregates are necessary. Additionally decorative fractions of special granulometries like calcite, marble, jurassic limestone, or mica are used, mainly for decorative plasters. To reduce the density of a dry mortar and to increase the insulation effect lightweight aggregates like perlite, vermiculite, glass foam, expanded clay, and pumice are used as additional aggregates. Because of their low density (typically 80 – 500 kg/m^3) only few per cent by weight is added to the mix. Dry mortars for decorative renders or tile grouts are often colored with pigments.

3.3. Additives

Without additives modern dry mortars would not exist and many technical properties could not be achieved. Relative to the mineral ingredients the content of additives typically lies between 0.1

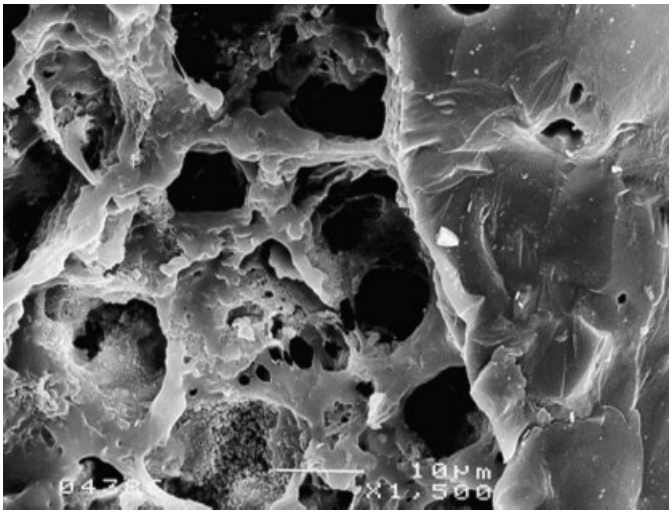

Figure 1. Scanning electron micrograph (× 1500; Wacker Polymer Systems) of the interface between a polymer-modified ceramic-tile adhesive (left) on a porcelain tile (right). The polymer films at the interface between the surface of the porcelain tile and the cementitious mortar can be seen clearly.

and 10 wt %. The additives are of organic or inorganic origin, often of polymeric nature. They can improve the mixing of the dry mortar with water, the properties of the wet mortar such as rheological behavior or the workability as well as the properties of the set mortar including the setting behavior.

3.3.1. Cellulose Ethers [12]

Cellulose ethers are used as thickening and water retaining agents in dry mortars. Cellulose ethers are of major importance as additives, even though their addition rate is very low (normally 0.02 – 0.7 %). Out of all additives cellulose ethers together with redispersible powders cause the largest range of effects in dry mortars. The mainly used cellulose ethers in dry mortars are methyl hydroxyethyl cellulose (MHEC) and methyl hydroxypropyl cellulose (MHPC). Together they have a market share in dry mortars of at least 90 %. Colloquially they are still called "methyl cellulose" or "MC", although pure methyl cellulose today has a very small market share. Other technically relevant cellulose ethers with a small share in the dry-mortar market are ethyl hydroxyethyl cellulose (EHEC) and hydroxyethyl cellulose (HEC). As (sodium) carboxymethyl cellulose is not stable in the presence of calcium ions, it is used only in few applications as a thickener.

The following sections cover relevant properties of MHEC and MHPC. As these properties are valid for both products, here they are simply referred to as MC.

Formation of Solutions and Wet Mortar Mixes; Viscosity Buildup. MC is water-soluble on a large temperature range. The most suitable MC's for dry mixes are powders whereby 20 – 60 wt % of the particles have a size smaller than 63 μm. The dry mix, in which MC particles are dispersed between binder and aggregate particles, avoids the formation of lumps which occurs only if powders are poured directly into water. Coarse MC products, normally classified as granular materials, are easy to dissolve in water without lump formation, but their slow dissolution makes them unfavorable for dry mortar mixes. For dry mortars with a neutral pH it should be considered that the granulometry is not the only parameter deciding the solution behavior of the MC. Some MC-grades are covered with a chemical crosslinking agent ("retarded solubility"), which causes particles to dissolve quickly only under alkaline conditions (e.g., coming from cement or hydrated lime). The alkalinity leads to an immediate breakdown of the crosslinking as well as to a fast dissolution

of the MC in the mortar. Originally MC grades with retarded solubility were not developed for dry mortar, however, they have eventuelly spread to the dry mortar market.

In pure solution and in a wet mortar, MC builds up a certain viscosity. The differences between low and high viscosities can be seen easily in a 2 % aqueous solution. Many MCs are specified by their viscosity measured at this concentration. The viscosities of such solutions varies between aqueous (viscosity up to a few hundred millipascals) up to jellylike (viscosity of several thousand millipascals). Different MC producers use different methods and apparatus to specify the viscosity of their MCs: the Haake Rotovisko, Höppler, Ubbelohde, and Brookfield methods are mainly in use. The results of viscosity measurements on the same sample can differ by up to several hundred per cent between two methods, and this should be borne in mind when comparing viscosities of MCs from different producers.

Stickiness and Workability. Stickiness is an expression mainly used in the leveling of plasters and renders. Here "stickiness" means the feeling of adhesion that the worker experiences to be between the leveling tool and the wall. High stickiness requires more force during leveling and results in a lower workability. Both properties can be influenced by the MC.

Water Retention. The water retention value (WRV) of a mineral plaster is the percentage of water that remains in the plaster after capillary dewatering by an absorbant substrate (DIN 18555, part 7) [13].

Cement- and gypsum-based mortars need water for setting, and this water must be retained in the mortar for a longer period of time. The thickness of thick-bed mortars widely used in the past (normally in the range of centimeters) protected the mortars from drying out too fast after coming into contact with water absorbing substrates, the sun, or other environmental conditions like wind, dry air, or a high ambient temperature. Today, wall materials with high capillary forces are used (e.g., aveated light weight concrete), and the thickness of mortar layers generally has decreased. MC is necessary to retain the water during the setting reaction. The high water retention of a modern dry mortar is mainly due to MC. Figures 2 and 3 show an example of the water retention of a wet mortar as a function of concentration and viscosity, respectively, of the added MC.

For the measurement of water retention see Section 5.

Water Demand and Yield. A mortar requires a certain consistency which is well known to skilled workers. They add as much water as is necessary to get the right consistency. The water demand depends on the ingredients and their addition rate in the formulation. MC is the main factor which influences the water demand. Mainly the viscosity, the addition rate, and the additional thickening effect of MC are the parameters. The water demand (water-solid-factor) influences also the yield of a mortar. It is measured in liters of wet mortar per 100 kg of dry mortar and is an important parameter for the efficiency of lightweight plasters. Some MC producers supply special MCs which garantees high yields in the mortar. This can help to reduce the amount of lightweight aggregates.

Other properties which can be influenced with MCs are open time and tensile adhesion strength of a tile adhesive as well as the slip resistance of a tile in the mortar (see Section 5), the rheological properties, plastification, and lubrication.

3.3.2. Other Additives

Starch Ethers. Mainly hydroxypropyl starches are added to plasters. Despite their low viscosities (mainly 100 – 500 mPa s in 2 % solution) they distinctly increase the viscosity of mortars when added to MC-containing mortars. Typical addition rates are between 0.01 and 0.04 % in cement-based renders and plasters, and 0.02 – 0.06 % in gypsum-based plasters. The water demand of the plaster is slightly increased, and this is associated with a slightly higher yield. The water retention of the mortar is not increased. With regard to the workability the sagging of the wet mortar from vertical walls and supports is reduced. At the optimum dosage the workabilities are improved.

Air-entraining agents act physically by entraining air micropores in the mortar. This leads

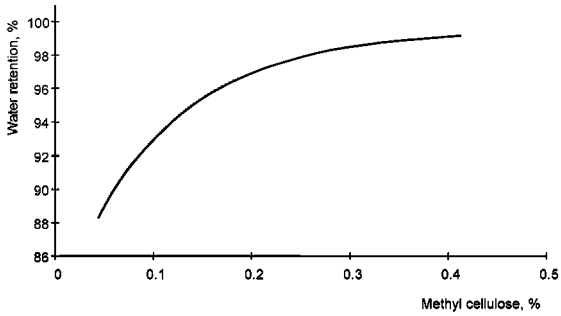

Figure 2. Water retention of a wet mortar as a function of the concentration of the added methyl cellulose (Walocel MKX 30 000 PP 01, Wolff Cellulosics, viscosity 30 000 mPa s, measured with a Haake Rotovisko)

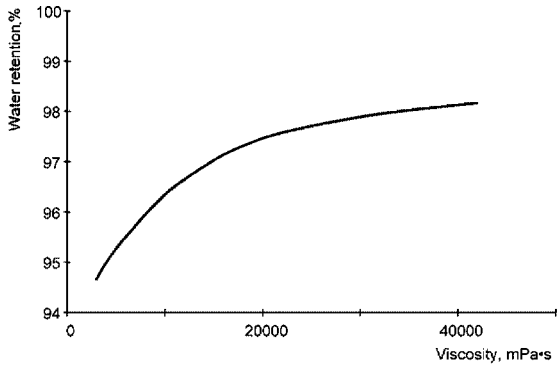

Figure 3. Water retention of a wet mortar as a function of the viscosity of methyl cellulose (2 % solution, 20 °C, shear rate: $D = 2.5\ s^{-1}$, measured with a Haake Rotovisko)

to a decreased wet mortar density, a better workability, and a higher wet mortar yield. The included air leads to better insulation against cold and heat, but also to lower strength. Air-entraining agents are based on powder form and mainly sodium salts of fatty acid sulfonates and sulfates. The addition rates in plasters and masonry mortars normally varies from 0.01 to 0.06 %. The optimum addition rates can be found by monitoring the air content of the mortar and its workability.

Accelerators. Accelerating systems are used in large amounts in cement-based systems to adjust the desired setting properties. In particular calcium formate (e.g., Mebofix®, Bayer, Leverkusen, Germany) or lithium carbonate (e.g., from Chemetall, Frankfurt, Germany) are used successfully. The addition rate is up to 0.7 % for calcium formate, and up to 0.2 % for lithium carbonate.

Retarders. The main application for retarders are gypsum plasters and gypsum-based joint fillers. Without retardation the setting of gypsum is too fast. Different retarders are used, mainly salts of fruit acids like tartaric or citric acid and of synthetic acids (e.g., Retardan® grades, from Tricosal, Illertissen, Germany). The typical dosage lies between 0.05 and 0.25 %.

Hydrophobic agents (water-repellent agents) prevent water from penetrating into the mortar, but the mortar remains still open for water vapor diffusion. The performance of hydrophobic agents can be measured by the capillary water absorption (DIN 52617) [14].

The main applications for hydrophobic agents are cement-based plasters for exterior application, mineral waterproofing slurries and tile grout. There are two groups of hydrophobic agents on the market: metal salts of fatty acids. (e.g., zinc stearate or sodium oleate, e.g., from Greven Fettchemie, Bad Münstereifel, Germany), and polymeric redispersible powders with hydrophobic properties (some Vinnapas® grades, Wacker Polymer Systems, Burghausen, Germany). The first group has the advantage of lower addition rate (0.1 – 1 %), the second one the advantage of significantly better durability, because hydrophobic redispersible powders are not washed out of the plaster by rain even after years. Additionally, the use of hydrophobic redispersible powders does not lead to wetting problems during mixing the dry mortar with water and it improves the adhesion of the cured mortar to the substrate (see Section 3.1.2).

Superplasticisers have a strong influence on the water demand of a mortar. A mortar containing superplasticisers requires less water than usual to get the same consitency. Consequently, an unchanged water demand leads to a lowering of the consistency. The effect of superplastification is explained by the model [15] that different surface charges of the cement particles lead to the inclusion of water when they agglomerate. By the adsorption of the superplasticiser the surfaces are discharged and water is set free. Depending on legal restrictions and technical advantages casein (many producers) or synthetic superplasticizers are used, e.g., on the basis of sulfonates of lignin, naphthalene, melamine – formaldehyde condensates, or polyether carboxylates. Examples of producers are SKW (Trostberg, Germany), Sika (Switzerland), and Perstorp (Sweden). Superplastizisers are mainly used in mortars which need very good self-leveling properties like self-leveling floor underlayments, screeds, and pourable flooring tile adhesives. The dosage normally lies in the range of 0.2 – 1 %.

Fibers can be distinguished into two groups: long fibers are mainly used for reinforcement of mortars. Short fibers (e.g., Arbocel and Lignocel grades from J. Rettenmaier & Söhne, Ellwangen-Holzmühle, Germany) are used in influence wet-mortar properties and water demand.

Defoamers reduce the air content in wet mortars (e.g., Agitan P products from Münzing Chemie, Heilbronn, Germany). Powder defoamers on a different chemical basis (mainly hydrocarbons, polyglycols or polysiloxanes on an inorganic carrier) are in use.

4. Production

Production, storage, transport, and quality control of dry-mix mortars are defined in DIN 18557 [16].

Modern dry-mix plants (Fig. 4) with a production capacity of typically 40 000 to 250 000 t/a mostly are built on a small area because the production line is oriented vertically, and the silos for the raw materials are placed above the mixing unit.

After appropriate quality control, the raw materials are transported by the receiving system to the different silos at the top of the plant. Consequently, the material flow is mainly gravitational, which saves on investment and running costs. The raw materials are transferred by gravity or by appropriate conveying systems (disk feeders, dosing screws, pneumatically) into the high-accuracy hopper-scales weighing system. Controlled by the fully automated electronic control system, the mixing unit is filled with all raw materials needed for each formulation for a specific dry-mix mortar. As mixing unit mostly special mixers are used which are suitable for the whole product range of dry-mix mortars (from fine-particle materials up to coarse dry mortars). Such mixing units, available in a wide variety of different sizes and designs, allow short batch cycle times and fast, homogeneous mixing. The temperature of the dry-mix materials should not exceed 50 °C during the whole mixing process to avoid deteriorating the thermoplastic and sensitive additives. After a short mixing process of about 3 – 10 min for highly efficient modern mixing units, the homogeneous dry-mix mortar is discharged into the intermediate finished-product storage silo. After quality control, the dry-mix mortar is discharged into transport silos or transferred to bagging and palletizing units, ready for the transport to the con-

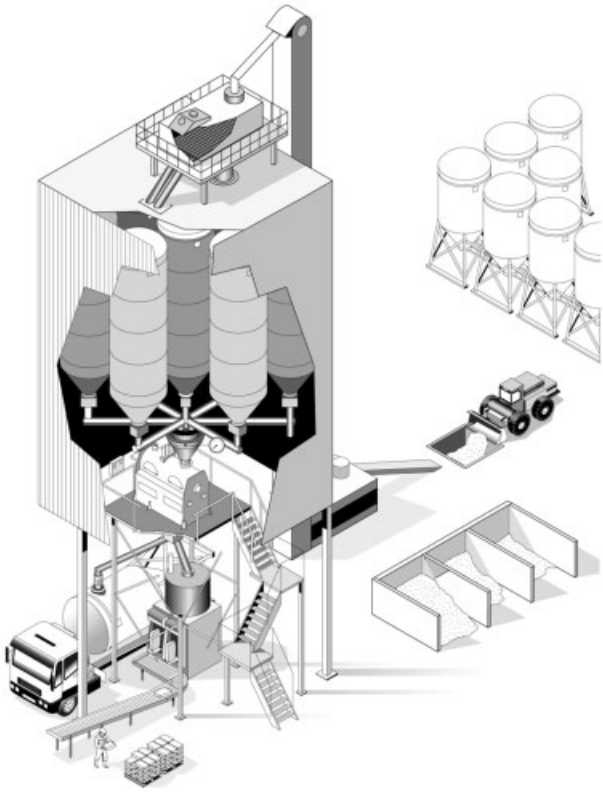

Figure 4. Schematic diagram of a dry-mix mortar plant (m-tec, Neuenburg, Germany)

struction site. Figure 5 shows a typical dry-mix mortar plant.

In the production of dry-mix mortars, the quality of all raw materials used, especially the bulk mineral materials, must be controlled according to the national standards (e.g., EN 196 for cements). If aggregates like silica quartz sands are not available in the appropriate quality, the dry-mix mortar plant must include installations for grinding, washing, drying, and classification into different sieve fractions. The residual humidity of all fillers should not be higher than 0.3 %, and the temperature of the sand after the drying process should not exceed 60 °C before being used. The sieve curve within the different filler fractions should be constant without large variations (e.g., by combining several subfractions).

The design, the size and the number of silos for all raw materials and the design of the whole mixing and packaging unit depends on the raw materials available and the number, types, and volume of different dry-mix mortars to be produced in the dry-mix mortar plant. For all gypsum-based products usually a separate production line is used to avoid contact or mixing of cement-based products with gypsum.

5. Testing

Consistency. Applying a mortar to a support requires to mix it with a certain amount of water. Skilled building workers instinctively mix the dry mortar with the amount of water which leads to the desired application consistency. A higher or lower amount of water leads to undesired properties. The required amount of water for mixing is described by the term w/s (water to solid ratio).

Figure 5. Typical dry-mix mortar plant (m-tec, Neuenburg, Germany)

During the development of a dry mortar the w/s values should be adjusted to a consistency which will probably be used also on building sites. If formulations with different addition rates of additives or different additives are compared, it is generally worthless to compare formulations with the same w/s, because nearly all additives change the consistency or the w/s slightly. To find a way of controlling the consistency and to compare different formulations, several testing procedures exist. The consistency of mortars, renders, and plasters is checked by the slump method, which is done in a similar way as in the concrete industry (→ Cement and Concrete, Chap. 2.4.1.). The standards DIN 1168, part 2 (only for gypsum-based products) [17], DIN EN 13279, part 2 [18], and DIN 18555, part 2 (for mortars with mineral binders) [13], define the slump to be "the diameter in millimeters of a pat made of a mixture of water and building plaster and formed by vertical blows after removal from the mold and jolting of the mixture" [13]. The consistency of more liquid products like self-leveling floor screeds or pourable flooring tile adhesive can be determined by the diameter of flow. This procedure does not require jolting. The consistency of a tile adhesive can be controlled by viscosity.

Water Retention. The standard DIN 18555, part 7 [13] describes the measurement of water retention. The water-absorbing wall is simulated by water-absorbing filter plates. The test is carried out on a plaster with a consistency as used in practice. The wet mortar is spread onto filter plates. After a certain time the water amount absorbed by the filter plates is weighed. Than the water retention is calculated in %. Different kinds of mortars require different water-retention values for good application properties.

Setting Time. The setting behavior of gypsum-based building plasters is determined by the Vicat needle test according to DIN EN 13279, part 2 and DIN 1168, part 2. [17], [18]: "The start of setting of building plasters containing additives is indicated by the time in minutes after which a Vicat needle, in the course of a penetration test into a plaster sample, comes to a standstill at a specified height." The time is calculated from the commencement of moistening of the building plaster with water. Coarser systems such as cement renders and plasters are less suitable for the Vicat testing procedure.

Air content. Freshly mixed mortars contain pores, one origin of them being air entraining agents, another being adsorbed air on the surface of the particles. DIN 18555 part 2 [13] describes a method to determine the total amount of air pores in a fresh mortar. It is an integral method which does not give any indication about the size or the distribution of the pores. "The air content of fresh mortar shall be measured by the pressure method using a preset test apparatus with a capacity of 1 dm^3. The test apparatus shall have a pressure chamber in which a defined pressure is produced. By opening an overflow valve a pressure balance is affected between the pressure chamber and the sample container (measuring vessel) which is filled with fresh mortar. The drop in pressure is a measure of the air content in the fresh mortar." [13]

Granulometric analysis mainly is done with screen sizing. The residues on the test sieves are weighed and expressed as a percentage of the initial weight. Plotting the diameter of the particles as a function of the percentage sieve passage gives the individual granulometric curve. The curve has a specific shape for each dry mortar. The coursness of dry mortars also varies (Fig. 6). The granulometric range of most dry mortars varies between 0.1 and 4 mm.

In some mortars a high wet mortar density is necessary to lead to high compressive strengths. In such cases the granulometric distribution of all mineral components should be optimized accordingly. To use a simple picture, the spaces between the bigger particles should be filled by smaller particles to give the best space filling. This best distribution of diameters of particles can be calculated and and plotted as a Fuller curve (Fig. 7).

Open Time. In the colloquial speech used on the building sites "open time" is a general expression for the length of time in which a freshly mixed mortar can be applied. The expression "open time" mainly is used for applying tile adhesives, sometimes also for plastering and rendering. A more precise definition of "open time" is given in the following three definitions: The former DIN 18156 (part 2) defines the "open time" to be the time between the application of the mortar layer and the formation of a skin. The skin formation can only be tested qualitatively with a wad of cotton. Another testing procedure defines the "open time" as the time after which the wetting of the back of the tile is at least 50 % of the square of the tile. Nowadays European laboratories are testing the open time mainly according to EN 1346 [19]. Here the open time (in minutes) is the largest time range after which the necessary tensile strength of a tile adhesive according to EN 12004 is reached. The tested times are 5, 10, 20, and 30 min.

Slip Resistance. A tile adhesive should have a good slip resistance mainly for the following two reasons: for tiling with heavy tiles like marble tiles, and second for countries where the tiling of a wall starts from the top (as in Germany) and not at the bottom. In the latter case the lower tile supports the upper tile and prevents it from slipping down.

The slip resistance is measured with a standard tile of size 10 × 10 cm and a weight of 200 g on a concrete slab according to EN 1308 [20].

Tensile adhesion strength is an important property of many dry mortars. For tile adhesives EN 1348 [21] specifies four conditions of storage: standard storage under standard climatic conditions for 28 d, after water immersion, after freeze – thaw cycle, and after heat aging storage.

6. Applications

6.1. Brick-Laying Mortars and Related Adhesives

Brick-laying mortars are used to join all kinds of bricks: red clay bricks with low water absorption, strongly absorbing sand-lime bricks, and aerated light weight concrete. Generally, bricks with low water retention applied in thick layers require mortars with low water retention, while smooth and even bricks with high water absorption require a mortar with high water retention applied in thin layers. Table 2 lists the different formulations.

DIN 18555, parts 1 – 9 [13] describes the testing of mortars with mineral binders. Most of them are valid for brick-laying mortars as well as for cement-based renders and gypsum plasters.

Table 2. Typical formulations of a brick-laying mortar and an aerated light weight concrete adhesive (in parts by weight)

Component	Brick-laying mortar	Aerated light weight concrete adhesive
Cement (e.g., CEM I 32.5R)	12 – 20	36
Hydrated lime	0 – 6	4
Limestone dust, 0 – 0.1 mm	10 – 20	
Quartz or limestone sand, 0 – 4 mm	60 – 80	
Quartz sand, 0 – 0.5 mm		60
Air-entraining agent	0.01 – 0.03	
Methyl cellulose, medium viscosity	0.02 – 0.04	
Methyl cellulose, high viscosity		0.3 – 0.4

6.2. Renders and Plasters

Plaster is defined in DIN 18550, part 1 as "a coating of plaster mortar or coating materials applied to walls and ceilings in one or more layers in a

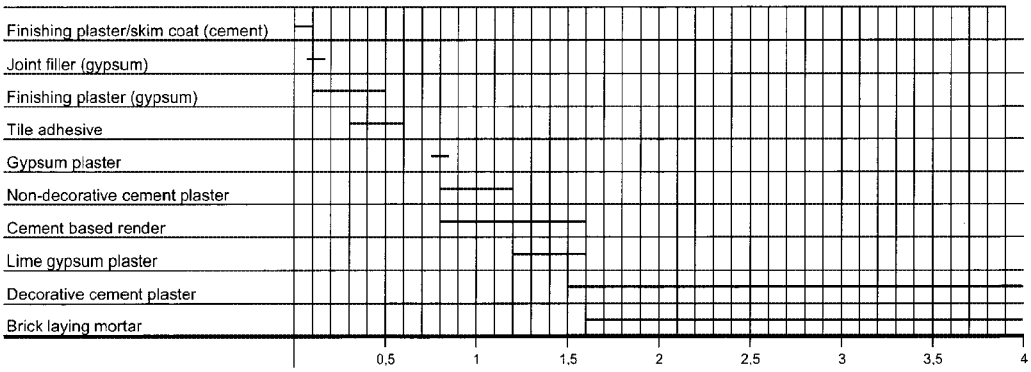

Figure 6. Maximum grain size of some European mortars (mm)

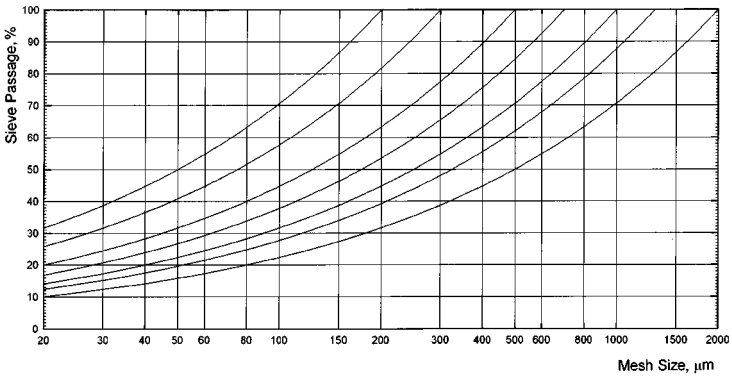

Figure 7. Reference curves calculated from the Fuller equation

specified thickness, which does not acquire its final characteristics until it has set on the building component" [22]. Different types of dry mortar render and plaster are classified on the basis of the type of binder used:

- Plaster and render mortar with mineral binders (cement, gypsum, and possibly hydrated lime)
- Decorative plaster mortar with cement, redispersible powder, and possibly hydrated lime as the binder (see Section 5)

Dry mortar render and plaster with other binders like potassium silicate, redispersible powder as the sole binder, hydrated lime as the sole binder, and clay are not covered here. Plasters assume a range of physical tasks, for instance, protection against weathering or chemical or mechanical actions. Weathering means the ingress of moisture or fluctuations in temperature. In addition plasters have to confer adequate protection against driving rain, but they are used also for bathrooms and other rooms where moisture occurs. Cement or lime – cement plasters are used to satisfy these requirements.

Render and plaster must have good water vapor permeability and be suitable for painting and hanging heavy papers. Mineral renders, typically applied in a single layer with a thickness of ca. 10 – 30 mm, also serve as a uniform and smooth substrate or load-bearing layer for subsequent finishing coating materials like ceramic tiles, paints, and decorative finishing coatings. Cement-based renders are used for exterior applications and wet rooms, whereas gypsum-based renders are used exclusively for interior walls.

Nowadays, in most European countries (not in the UK) machine-applied render/plaster is much more common than manually applied render or plaster. Accordingly, the render/plaster for machine application must fulfill the additional requirements of a very rapid development of consistency and high water retention. Both parameters are steered by choosing the appropriate grade of methyl cellulose. Next to the trend to machine application, there is a tendency to use more lightweight plasters.

Cement and Lime – Cement Renders. The main requirements for cement and lime – cement renders are high non-sag properties during application, easy workability, low stickiness to the leveling tool, and crack-free setting. To fullfill these requirements, the formulation requires the optimum concerning granulometry and the addition of additives like methyl cellulose, starch ethers, and air-entraining agents. Table 3 lists typical formulations of a lime cement render and of a lightweight render based on cement.

Table 3. Typical formulations of a standard lime – cement render and a lightweight cement render (in parts by weight)

Component	Lime – cement render	Lightweight cement render
Portland cement CEM I 32.5R	8 – 12	18 – 25
Hydrated lime	6 – 8	0 – 5
Quartz sand, 0.2 – 0.8 mm	80 – 85	
Limestone sand		60 – 75
Limestone dust		5 – 7
Expanded polystyrene		1 – 2
Starch ether		0.01 – 0.02
Hydrophobic agent	0.15 – 0.25	0.1 – 0.2
Air-entraining agent	0.015 – 0.03	0.03 – 0.05
Methyl cellulose, viscosity 15 000 mPa s	0.08 – 0.12	0.1 – 0.12

Lightweight renders for thermal insulation are specified in DIN 18550, parts 3 and 4 [22].

Gypsum Plasters (→ Calcium Sulfate). As the working steps necessary for a gypsum plaster are more complicated and time-consuming than for a cement render, the typical formulations are also more complicated. The main focus is always on the correct adjustment of the setting retardation, which enables the worker to finish all steps of the work before the surface sets or becomes dry. As the working traditions are different in many countries, the local formulations always have to be adjusted to the different working procedures. A main difference in working concerns the final working steps after the second leveling. In central Europe the surface is wetted with water and a cream is rubbed out of the surface with a sponge, followed by the smoothening with a knife, which results in extremely smooth surfaces. In southern Europe and Great Britain, watering of the surface is avoided, and this leads to a coarser surface. To obtain the desired smoothness, the surface is plastered with a fine-grained top coat.

Lightweight gypsum plasters containing perlite or vermiculite are very common. Especially in these formulations methyl cellulose plays an important role in increasing the yield and saving on the cost of the lightweight aggregates. In a lightweight gypsum plaster for machine application the right MC allows 100 – 200 L of perlite to be saved per tonne of plaster for the same coverage compared to a MC without any additional thickening effect. Such an MC will yield to an approximately 10 % higher coverage of the plaster mortar.

The true yield of a machine-applied lightweight plaster can be measured only in a plaster spaying trial. Typical German gypsum plasters have yields in the range of 80 L/100 kg of dry mortar for a gypsum – lime plaster, and up to 120 L/ 100 kg of dry mortar for machine-applied gypsum lightweight plaster.

Relevant standards are draft DIN EN 13279, parts 1 and 2 [18] (formerly DIN 1168, parts 1 and 2).

Table 4 shows typical formulations of a gypsum plaster, a gypsum – lime plaster, and a lightweight gypsum plaster.

Textured Mineral Finishing Plasters. Mineral finishing plasters, stuccos, putties, and skim coats are part of a coating system and are applied usually on a thick-layer base coat (render) or on other types of uniform and smooth surfaces (concrete, plasterboard, etc.). For exterior walls, mainly lime – cement plasters are used as a decorative finishing material. Finishing plasters must provide the facade with an optically attractive surface finish and have to fulfill physical functions such as a durable protection of the covered wall areas and undercoats against dampness and weathering.

Table 4. Typical formulations of a gypsum plaster, a gypsum – lime plaster, and a lightweight gypsum plaster (in parts per weight)

Component	Gypsum plaster	Gypsum – lime plaster	Gypsum lightweight plaster
Gypsum	74 – 98	40 – 50	70 – 100
Limestone		20 – 35	
Hydrated lime	1.5 – 5	15 – 20	2 – 5
Perlite		0.3 – 0.8	3 – 5
Starch ether	0.01 – 0.04	0.01 – 0.03	0.01 – 0.05
Air-entraining agent	0.015 – 0.03	0.015 – 0.03	0.01 – 0.03
Setting retarder	0.025 – 0.05	0.025 – 0.04	0.025 – 0.04
Methyl cellulose, viscosity 30 000 mPa s	0.16 – 0.23	0.16 – 0.23	
Methyl cellulose for high yield			0.2 – 0.24

To meet their required physical functions over long periods, mineral finishing coating materials for exterior application must have a good adhesion to the substrate, a low water absorption and a high water-repellent effect (low coefficient of water-absorption), good drying characteristics (good water-vapor permeability), and low susceptibility to cracking; the modulus of elasticity of the mineral coating should be lower than the modulus of elasticity of the layers below it.

These requirements are met in an exemplary manner by mineral-textured finishing plasters produced as factory made dry-mix mortars. The term "finishing plaster" nowadays refers to facade-coating materials pigmented in white or light pastel colors which are used as the final surface finish without any additional painting being necessary. Mineral finishing plasters are composed of hydrated lime and cement as mineral binders, aggregates (fillers), pigments, and additives such as cellulose ethers and starch ethers to improve the water retention and processing. If required, other additives such as air entrainers, water-repellent agents, retarders, and fibers are added. The ratio of cement and hydrated lime can be varied according to the different requirements; the higher the cement content the higher the compressive strength, the toughness, and the water resistance, but there is also a higher risk of crack formation due to the brittleness and shrinkage of the mortar. The higher the content of hydrated lime, the better the workability, and the lower the compressive strength. Carbonate fillers such as marble or limestone can be used additionally or even exclusively instead of silica sands. The technical performance of the finishing plaster can be improved significantly by the addition of organic polymer binders in the form of redispersible powders. The different structures of the hardened mortar originate from a different particle size distribution of the fillers used in the formulation (structure-giving large grains) and from the application method (application by spraying, brush, trowel, roller, etc.). Mineral finishing plaster is applied manually but nowadays increasingly by machine spraying.

The standard DIN 18550 [22] defines the technical requirements and the application of mineral base renders as well as of the finishing plaster coating systems.

6.3. Tile Adhesives

Ceramic wall cladding for building interiors and exteriors is by no mean a new invention. The first ceramic wall cladding materials in form of fresco and mosaic tiles were produced ca. 3500 years ago in Egypt, Persia, and China. Tiles provide an esthetically attractive decorative surface in combination with important functional benefits, being water resistant, tough, long-lasting, hygienic, and easy to clean. For these reasons tiles are important floor and wall covering materials in the construction industry. In 1999 ca. 4.5×10^9 m^2 were produced and laid worldwide. The most important market was Asia (1.9×10^9 m^2) followed by Europe (10^9 m^2) and America (0.9×10^9 m^2).

Mainly earthenware tiles for indoor applications (non-frost-resistant porous tiles with a water absorption of ca. 25 % according to the EN 87 [23] and frost-resistant stoneware tiles (non-porous tiles with a water absorption of ca. 1 %) for indoor and outdoor applications are used. Different types of natural stones are installed for indoor and outdoor applications. Nowadays,

the hard-pressed fully vitrified tiles (porcelain tiles) with an extremely low water absorption (< 0.1 %) and an excellent scratch, wear, and weather resistance, as well as glass tiles, are being increasingly used, mainly for outdoor applications and for floors. In addition there is a strong trend to use tiles in larger sizes (large-formatted tile, up to 40×40 cm).

Ceramic tiles and natural stones previously were exclusively laid with the thick-bed mortar technique based on job-site-mixed mortars. In this method, sand and cement are mixed together on the job site to produce a simple cement mortar with a cement/sand ratio of ca. 1/4 to 1/5. After applying (buttering) this mortar on the reverse side of the water-(pre)soaked or prewetted tile in a thickness of 15 – 30 mm, the tile buttered with the mortar is pressed to the prewetted surface to be tiled. The tiles are tapped to ensure uniform flatness of the tile surface, and thus a final mortar bed of 10 – 25 mm is obtained. This procedure not only causes a compaction of the mortar, but leads in addition to the migration of the fine cement particles into the porous reverse side of the tiles and into the porous substrate. In this way a mechanical fixation or anchoring of the tile in the mortar bed is assured, as well as anchoring of the mortar on the substrate after curing of the cement. Because this simple type of mortar has no slip resistance, tiling must be started at the bottom, and spacers must be used to obtain regular joints between the tiles.

However this method is a very time-, cost- and, material-consuming process which requires experienced craftsmen. They have to decide whether the substrate and the tiles are suitable for using this method, a certain soaking time of the tiles is required depending on their porosity, the mortar must be mixed in the correct ratio and consistency, and the right amount of mortar has to be applied on the reverse side of the tile before it is laid. More significantly, there are a lot of technical restrictions in using this technique in the modern building industry. For example, only porous and relatively small format tiles can be laid on porous, solid, and strong mineral surfaces (backgrounds) with this method due to the need for mechanical anchoring of the mortar.

Therefore, the thin-bed mortar technique today has replaced the thick-bed mortar method in most industrialized countries. After the polymer-modified prepacked dry-mix mortar has been gauged with water, it can be applied to a large area of the surface to be tiled with a notched trowel (floating technique) to give a ribbed mortar bed of uniform thickness. Due to the good water retention capacity of the thin-bed mortar (effect of cellulose ether), neither the tiles nor the substrate (background) have to be presoaked or prewetted. The tiles are then pressed into the mortar bed with a slightly twisting action. If the adhesive mortar is formulated in the appropriate way by using adequate additives, tiles that have just been laid in the fresh mortar bed will not slip. Thus, the insertion of spacers between the tiles is not necessary and tiling can be carried out from the top to the bottom. This technique creates an adhesive mortar bed of ca. 2 – 4 mm thickness, depending on the dimensions of the trowel notches used (usually $6 \times 6 \times 6$ mm, depending on the size of the tiles and the flatness of the substrate). The thin-bed mortar technique is thus more cost-effective than its thick-bed counterpart. It also uses less material, can be applied more universally, and its execution is simpler, faster, and safer. This is even true in cases where the background is uneven and has to be coated or levelled with an equalization mortar first.

With this thin-bed technique (Fig. 8) using specially designed prepacked dry-mix adhesive mortars, all technical demands of modern building industry using different types of backgrounds and covering materials under different and extreme climatic conditions can be satisfied. Today there is wide range of ceramic-tile adhesives available, depending on the substrate to be tiled and the tiles being used: standard and flexible, normal- and fast-setting, as well as special adhesives such as white mortars for fixing natural stones, waterproofing adhesives, flow-bed mortars for floor tiling, gypsum-based adhesives, and high-flexibility mortars for fresh screeds.

Figure 8. Application of a thin-bed ceramic tile adhesive (spreading the adhesive mortar with a notched trowel on a wall to be covered with ceramic tiles); picture: Wacker Polymer Systems

Dry-mix mortars for thin-bed tiling must fulfill technical requirements such as good workability characteristics, good water-retention capability, long open time and adjustability time even at high temperatures, and good non-sag properties. After curing, the cement-based adhesive mortar must provide a good adhesive and cohesive bond strength between all types of covering materials (e.g., natural stones and all kind of ceramic tiles) and various backgrounds (e.g., concrete surfaces, brickwork, lime – cement renders and base coats, gypsum, wood, old tile surfaces, gypsum wallboards, aerated lightweight concrete, particle boards, etc.). This must be guaranteed even after exposure to frost, dampness, and even permanent immersion in water. In addition to good adhesion, the thin-bed adhesive mortar must have sufficient flexibility to absorb and reduce possible tensions between the background and the tiles caused by different thermal expansion coefficients of the covering materials and the substrates as well as possible movements of the background.

These overall characteristics of a thin-bed adhesive mortar can only be achieved by prepacked polymer-modified cementitious dry-mix mortars containing cellulose ethers as additives and redispersible powders as a polymer binding agent. Typical basic formulations for a standard and a high quality flexible ceramic tile adhesive are given in Table 5.

Table 5. Typical formulations for ceramic-tile adhesives

Adhesive type∗	A	B
Portland cement (OPC)	45	35
Silica sand (0.05 – 0.5 mm)	53.1 – 51.6	59.6 – 57.6
Cellulose ether (viscosity ca 40 000 mPa s)	0.4	0.3
Redispersible powder	0 – 4	5 – 10
Additives (if required for special performance)	(0 – 5)	(0 – 5)

∗ Type A = standard polymer-modified ceramic-tile adhesive, Type B = flexible, high-quality, polymer-modified ceramic-tile adhesive.

The most important technical requirements in Europe for a ceramic tile adhesive (besides workability) and the test methods are summarized in Table 6.

6.4. Tile Grouts

Tile grouts are used to fill the joints between tiles or natural stones laid on walls or floors. Cement-based tile grouts with appropriate formulations are suitable for interior and for exterior application. In combination with the tiles they must provide an optically attractive suface and must perform physical functions. The tile grout must be capable of reducing stresses within the whole wall or floor covering material, it must protect the materials and layers under the tiled surface against mechanical damage and the negative influences of water penetrating into the whole construction. Thus a cementitious grout must have good adhesion to the edges of the tiles, low shrinkage, sufficient deformability or flexibility, high abrasion resistance, good toughness and cohesion, low water absorption, and an excellent workability (low stickiness of the wet mortar). According to their application, two main types of tile grouts can be classified (Table 7).

For white and colored tile grouts, white cement is used as a mineral binder. Alkali-resistant pigments such as iron oxides should be used for colored grouts. Appropriate carbonate fillers can be used additionally or even exclusively instead of silica sand as a filler. To reduce the risk of efflorescence, the use of hydrophobing agents is very helpful, and microsilica or trass can additionally be added. The main technical requirements for tile grouts are summarized in Table 8.

Table 6. Summary of European standards for ceramic-tile adhesives

Test	Standard	Requirements
Definitions and terminology	EN 1322	
Slip resistance	EN 1308	< 0.5 mm
Wetting capability	EN 1347	
Tensile adhesion strength	EN 1348	Minimum requirement: > 0.5 N/mm^2 for all storage conditions
Storage conditions: standard conditions: 28 d sc∗ water immersion: 7 d sc + 21 d in water heat aging: 14 d sc + 14 d 70 °C +1 d sc freeze-thaw: 7 d sc + 21 d in water + 25 cyclesfreeze-thaw		Additional requirement: > 1.0 N/mm^2 for all storage conditions
Heat-aging test: 14 d sc + 14 d 70 °C + 1 d sc		
Freeze – thaw resistance: 7 d sc + 21 d in water + 25 cycles feeze – thaw)		
Open time	EN 1346	> 20 min; or > 30 min
Deformability/flexibility with transverse deformation test	EN 12002	
Adhesives for tiles – definitions and specifications	EN 12004	

∗ sc: 23 °C/50 % R.H.

Table 7. Typical formulations for tile grouts

Tile grout type∗	A	B
Portland cement	25 – 30	20 – 25
High-alumina cement (HAC)	0 – 10	0 – 10
Pigments (TiO_2; iron oxides)		0 – 5
Filler (silica sand and/or carbonate filler)	75 – 56.9	79 – 51.9
Cellulose ether	0 – 0.1	0 – 0.1
Redispersible powder	0 – 2	1 – 5
Additives for workability	0 – 1	0 – 3

∗ A: standard gray tile grout for interior and exterior use, B: high-quality, pigmented, smooth-surface tile grout for interior and exterior use.

6.5. Exterior Thermal Insulation Composite Systems

External thermal insulation composite systems (ETICS) or external insulation and finishing systems (EIFS) were developed in Europe in the early 1970s. The first oil crisis in Germany in 1973, together with financial support from the government for homeowners, helped tremendously to promote the system. From 1973 to 1993 ca. 300×10^6 m^2 of facades in the Federal Republic of Germany were equipped with ETICS [24], and more than 18×10^9 L of heating oil was thus saved. Apart from saving energy, emissions of pollutants and CO_2 are reduced.

ETICS also significantly increases the comfort while living in a building due to a more constant temperature and humidity over the seasons. It reduces significantly building damages by decreasing the temperature variations in the external wall construction and combined with less condensation of dampness allows to reduce the overall building costs. Therefore, thermal insulation is not only a sensible investment in new buildings, but also a good renovation measure, especially if the facade has to be renovated anyway.

The first ETICS, introduced in the 1970s, consisted of a dispersion-bound paste adhesive to which cement was added at the construction site. This compound was used to fix expanded polystyrene panels to the walls to be insulated. The same product and procedure was used to embed the fiber-glass mesh that serves as a reinforcement in the reinforcement layer or base coat on top of the fixed thermal insulation panel. In earlier years mainly dispersion-bound synthetic resin plasters were applied as a finishing coat after the base coat had been primed.

However, in practice many mistakes occurred when using this system, in particular during the mixing of the dispersion-bound paste adhesive with cement at the construction site. With this system it is not possible to always produce a homogeneous mixture with a constant polymer to cement ratio, and adhesives and base coats with different compositions are thus obtained. This results in inadequate product characteristics and possible damage to the facades due to insufficient adhesive bond strength or flexibility. This argument also apply to the two-component systems, in which a water-based dispersion is added to dry-mix mortars at the construction site. For

Table 8. Important requirements, test methods, and standards for cement based tile grout mortars

	Standard	Requirements
Grouts for tiles – definitions and specifications	prEN 13888	
Determination of water absorption	prEN 12808-5	water absorption (WA) determined with $16 \times 4 \times 4$ cm prisms; requirements: fundamental characteristic: WA after 30 min < 5 g; WA after 240 min < 10 g additional characteristic: WA after 30 min < 2 g; WA after 240 min < 5 g
Determination of resistance to abrasion	prEN 12808-2	abrasion tester according to EN102; requirements: fundamental characteristic: volume removed during test $< 2000\,mm^3$ additional characteristic: volume removed during test $< 1000\,mm^3$
Determination of shrinkage	prEN 12808-4	shrinkage determined with $16 \times 4 \times 4$ cm prisms after 28 d sc; requirement: < 2 mm/m
Determination of flexural and compressive strength	prEN 12808-3	flexural (FS) and compressive strength (CS) determined with $16 \times 4 \times 4$ cm prisms after 28 d sc and freeze – thaw cycles; requirements: FS $> 3.5\,N/mm^2$; CS $> 15\,N/mm^2$

these reasons, togther with the above-mentioned principle disadvantages of two-pack systems, today the polymer-modified cement based dry-mix mortars have almost completely substituted all other systems for ETICS.

Classic ETICS consist of the following components:

- Adhesive for fixing the thermal insulation panels, mainly expanded polystyrene, on the wall (depending on instructions and technical guidelines, additional mechanical fixing with special dowels may be necessary).
- Base-coat for embedding the reinforcement mesh on top of the thermal insulation panels (mechanical protection of the thermal insulation panels).
- Mineral finishing coating (mainly cementitious plasters or stucco with different structures).

A typical formulation (in wt %) for the adhesive and base-coat mortar is given in the following:

Portland cement	20 – 30
Filler (silica sand and/or carbonate filler, 0.05 – 0.5 mm)	64.7 – 75.9
Cellulose ether	0.1 – 0.3
Redispersible powder	4 – 5
Additives	0 – 3

The combination of individual components of an ETICS must always be seen as a whole. Thus the adhesive, thermal insulation panels, embedding mortar, fiber mesh, primer (or keycoat if necessary), dowels (if an additional mechanical fixing is required), and the finishing coat must be carefully adjusted to each other and must be tested and approved together as a system. Therefore, it is not permitted to combine individual components from different sources to create an ETICS which has not been tested and approved as a whole system.

Figure 9 shows a typical composition of a ETICS.

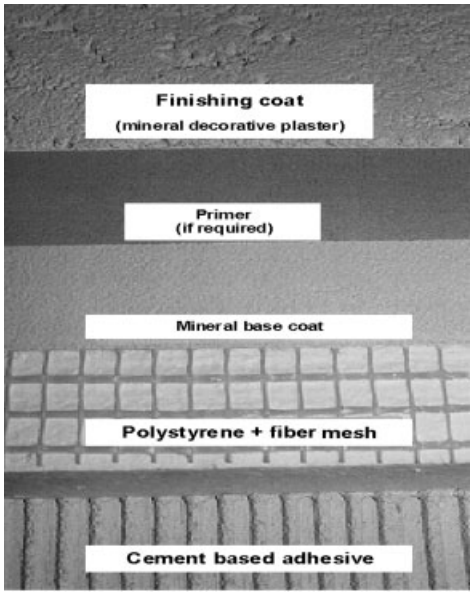

Figure 9. Typical components of a ETICS (Wacker Polymer Systems)

To meet the technical requirements, ca. 4 – 5 % of organic binder in form of a redis-

persible powder is used for the adhesive mortar and the base coat. This means that apart from the required good adhesion of the cementitious mortar on the substrate and on the insulating polystyrene panels, a sufficient deformability (flexibility) and a high degree of impact resistance of the mortars are achieved as well.

A large choice of different materials and systems are available nowadays as finishing coats for the ETICS. Apart from the classic dispersion-bound plasters, available in many structural variations and colors, mainly mineral cement – lime plasters or stuccoes, silicate plasters, and to a certain extent silicone resin plasters are used.

Various countries have differing regulations and specifications for ETICS. At present a European standard is being prepared for external thermal insulation composite systems [draft from European Organisation for Technical Approvals (EOTA)]. It will describe all specifications for the individual components of the system and the overall composite system itself (e.g., physical and construction specifications as well as techical requirements like water absorption, behavior under hydrothermal stress, inflammability, water-vapor permeability, adhesive strength, impact resistance, etc.).

Worldwide, new thermal insulation regulations (e.g., WschVO in Germany [25]) furthermore establish minimum requirements for the K-value (thermal insulation value) of the external walls of new buildings. For concrete and bricks walls these binding minimum requirements can often only be met by the additional application of an ETICS, and this will lead to a further growth in thermal insultation systems. Even in countries with an high average temperature, thermal insulation composite systems are becoming popular for saving energy for air-conditioning.

6.6. Powder Paints

The use of liquid dispersion paints is well known and has proven itself in practice since their invention in the 1950s, and today this is the leading technology for interior and exterior paint application. As a sole binding agent for his type of products aqueous dispersions based on styrene – acrylate, vinyl acetate ethylene, or other copolymers are used.

In DIN 55945 surface coating materials such as laquers and paints are defined, classified, and characterized. Paints according to this standard are compounds in liquid and pasty form, based on solvents, binding agents, pigments, fillers, and additives, but including powder-form coating materials as well. Paints can be classified as:

- Solvent-based coating materials
- Water-based coating materials (laquers, paints)
- Spray-applied powder coatings
- Paints and coating materials in powder form (powder paints) to be mixed with water before application

Powder paints were known for many decades before the invention of the liquid dispersion paints. In the early days of the powder paints the binder was either of organic origin (e.g., distemper paint based on glue) or of inorganic origin (e.g., mineral paints based on the mineral binders hydrated lime and/or cement). For a long time the mineral binder was modified with organic binding agents (casein, cellulose ether, starch ether) to improve the technical characteristics of the mineral coating material.

Natural organic binders have since been replaced with much more efficient synthetic polymer binders in the form of redispersible powders, and this has again led to an increasing importance of powder paints. Redispersible powder as organic binder is used as the sole binder in powder dispersion paints and for the modification of mineral powder paints based on cement, hydrated lime, or water glass (silicate powder paints) to improve their technical performance and durability.

Powder dispersion paints were first used on building facades in Austria in the 1960s and later in Germany. They provide almost the same characteristics as liquid dispersion paints, but offer all the advantages of a powder system. Powder dispersion paints are used for interior and exterior applications. The most important difference in formulation is the content of polymer binder. The higher the polymer content, the better the adhesion and the higher the wet-scrub resistance of the paint.

Like typical dry-mix mortars, powder paints are supplied in paper bags or in silos to the job site and are gauged with water befor use and are applied by brush, roller or spraying.

Table 9. Important test methods and standards for (powder) paints

	Standard	Comments
Definitions	DIN 55945	definitions, composition, types of painting materials,
General requirements	DIN 53778	general requirements concerning workability, capability of dilution, tinting, compatibility with pigments, minimum temperature of application, recoatability, gloss, hiding power
Water absorption	EN ISO 15148	coefficient of water absorption $W < 0.5\,\mathrm{kg\,m^2\,h^{0.5}}$ and $W \cdot S_\mathrm{d} < 0.2$ $\mathrm{kg\,m\,h^{0.5}}$ for exterior application (EN ISO 15148 substitutes DIN 52617)
Water-vapor permeability	EN ISO 12572	diffusion-equivalent air-layer thickness $S_\mathrm{d} < 2\,\mathrm{m}$ (EN ISO 12572 substitutes DIN 52615)
Wet-scrub resistance	DIN 53778	$>$ 1500 cycles for indoor application, $>$ 5000 cycles for outdoor application and for scrub-resistant paints

Table 9 summarizes the most important standards and technical requirements for powder paints.

6.7. Cementitious Waterproofing Sealing Slurries

Water in liquid or in vapor form is the most destructive weathering element for building materials such as concrete, masonry, and natural stones.

Traditional sealing and waterproofing systems (e.g., according to DIN 18195) include bituminous materials, plastic waterproofing foils, and metal tapes for interior and exterior applications. Different types of materials can be used to seal and protect the surface of buildings or their structural components against the intrusion of dampness and water. Nowadays products for this purpose are based on reactive resins like epoxy and/or polyurethane, dispersions (paintable waterproofing membranes), and mineral binders like cement, which are known as waterproofing membranes or sealing slurries.

Cementitious waterproofing membranes have been successfully used for more than 40 years in Europe. The structures to be protected were either exposed to periodic or long-term wetting (surface water, seepage water), low hydrostatic pressure (soil dampness), or in combination with appropriate engineering even high hydrostatic pressure. Cementitious membranes (slurries) are used to waterproof wet rooms and water tanks, and due to their excellent weathering resistance also for exterior surface protection. Further typical applications are the sealing and waterproofing of basement walls, of swimming pools, as well as walls and floors in bathrooms, and on balconies and terraces as a waterproofing layer to be tiled over. Some of the main advantages of cement-based waterproofing membranes are their excellent resistance to water (even if exposed permanently), excellent long-term weathering resistance, good scratch resistance, good load-carrying capacity and much higher water-vapor permeability than most of the other systems.

Figure 10 shows a typical application of a cementitious waterproofing sealing slurry by brush.

Figure 10. Typical application of a cementitious waterproofing sealing slurry by brush (Wacker Polymer Systems)

Cement-based waterproofing slurries are easy to use, nontoxic, provide a fully bound and monolithic surface without joints, and can be

easily applied on substrates with complex surface shapes. In contrast to other systems, cementitious waterproofing slurries can even be used on damp and wet mineral surfaces. Their physical properties are also less temperature-dependent compared to bitumen-based materials. So far there are no standards available for these alternative sealing products, but national guidelines of mortar associations are well accepted in the building industry [26].

Simple, non-polymer-modified cement-based slurries are still used for protection against surface water, but they are not suitable for sealing against water under hydrostatic pressure. To improve the poor adhesion, the poor water tightness, and the extremely low deformability and flexibility of these unmodified systems, polymers are added in form of liquid dispersions on the job site or as a redispersible powder already mixed with the dry-mix mortar. Special additives in these dry mix mortars include water-retention agents, thickening agents, and rheological additives.

Today standard or rigid mineral waterproofing slurries are polymer-modified, prepacked dry-mix mortars with a high cement and relatively low polymer (redispersible powder) content. They are used for mineral substrates which are dimensionally stable, sound, and solid without risk of crack formation, movement, or dimensional changes such as shrinkage. Developments in the late 1970s led in Europe to flexible waterproofing slurries, which to a certain extent are capable of bridging small cracks (up to ca. 1 mm) in the substrate. The flexibility of such products strongly depends on the polymer to cement ratio and on the flexibility of the polymer itself. Flexible and highly flexible waterproofing cementitious slurries are used on substrates that are still subject to shrinkage, vibration, movement, stress, and crack formation, and on substrates that are difficult to coat, such as wood, steel, aerated lightweight blocks, and gypsum. Due to their high polymer content (up to 40 wt % of the total formulation), they have a high diffusion resistance and are chemically resistant to chloride, sulfate ions, carbon dioxide, or other aggressive substances. In central Europe flexible cementitious waterproofing sealing slurries represent the last domain for two-component (two-pack) systems, which are based on prepacked dry-mix mortar and liquid admixes. However, even in these high-polymer products, there is a strong trend towards one-pack systems in form of polymer-modified dry-mix mortars.

6.8. Self-Leveling Underlayments and Screeds

Self-leveling underlayments (SLUs) are from a technical perspective probably the most complex field for dry-mix mortars. On a given uneven substrate (i.e., screed or surface to be refurbished), self-leveling mortars have to provide a suitable, smooth and solid substrate for applying all kinds of flooring materials like carpets, wooden parquet, PVC, tiles, etc. Self-leveling underlayments should be applicable in an easy and efficient manner, even for large areas. Therefore, the SLU material has to have very good flow characteristics, and self-leveling and self-smoothing properties. In addition, it should set and dry quickly so that the flooring material can be applied on top of the hardened mortar after only a few hours. The SLU mortar should adhere to all kinds of substrates and exhibit low shrinkage, high compressive strength, and good abrasion resistance.

The technical requirements for SLUs range from very simple to highly sophisticated products. They vary in thickness from a very thin layer of 1 – 10 mm (feather finish, self-leveling/troweling mortars and underlays), up to 60 mm for self-leveling screeds, which are always applied by machines (mixing and pumping in one set up).

The screeds are used as a load wearing base in a thickness from 30 to 80 mm, and are thus high-volume-application mortars. These mortars can be based on cement or gypsum (anhydrite) as mineral binder, the latter providing underlayments with a high dimensional stability (low or even no shrinkage). The set time ("walk-over time") of these materials varies from normal/regular setting to fast-setting products. Normally this depends on the requirements of a specific job, that is, the flooring material can be laid on the SLU in a certain time frame. The shorter the setting and drying time, and the thicker the mortar is applied, the more complicated and expensive the formulation becomes. Self-leveling compounds for underlayments and screeds are based on special hydraulic binders like Portland

cement (OPC), high-alumina cement (HAC), and gypsum (anhydrite), to achieve fast curing and drying by avoiding excessive shrinkage or expansion.

So far there are no standards on self-leveling underlayments (SLU) in Europe. However, the technique and its application have been well known for many years, and exclusively polymer-modified prepacked dry-mix mortars are used for this application.

6.9. Patching and Repair Mortars

Concrete is a very versatile, long-lasting, and durable building and construction material if it is applied according to the state of the art. In the past, and even today, unfortunately, repeated disregard of the fundamental principles of concrete and structural concrete application still leads to severe and serious damage in the building industry. The cost of the repair of concrete structures has dramatically increased over the last 30 years in all industrial countries. In Germany approximately 20 % of the cost of the volume of structural concrete work is attributed to the repair and maintenance of existing buildings and structures. The degradation of structural concrete is caused by corrosion of the steel reinforcement due to carbonation of concrete caused by atmospheric carbon dioxide and other aggressive media (e.g., SO_2, acid rain). The corroded steel reinforcement increases in volume and splits off the overlying concrete, thus destroying the structure.

In the construction industry concrete repair work can be classified into two types:

1) Repair of concrete which does not contain steel reinforcement and which does not have load-bearing functions. It is normally carried out for aesthetic reasons (cosmetic repair work) only, with patching mortars or patching compounds.
2) Repair and reconstruction of damaged reinforced and load-bearing concrete structures to maintain and reconstitute their structural stability and function. This is carried out in stages with different kind of mortars, which are part of a "concrete rehabilitation system" (typical applications: repair work and rehabilitation of bridges, parking decks, tunnels, etc).

Patching mortars for reprofiling and cosmetic repair are based on dry-mix mortars. Usually cement-based mortars are used for indoor and outdoor applications, whereas gypsum-based products are only used for some specific indoor applications (cosmetic repair), i.e., for filling small holes, voids, cracks and cavities to restore the original dimensions.

To guarantee the durable and reliable repair of structural concrete, concrete rehabilitation systems must restore the corrosion protection of the steel reinforcement (alkaline environment), reprofile the concrete structure, restore its load-bearing functions, and finally restore the protection and thus durability of the whole construction (protection against weathering and environmental damage caused by CO_2, SO_2, Cl_2, deicing salts, etc).

For the rehabilitation of structural concrete three types of mortars are used today: cement-based concrete mortars (CC; mostly applied as a shotcrete mortar), polymer (modified) cement concrete (PCC) as prepacked dry-mix mortars, and epoxy cement concrete (ECC).

Today mainly PCC mortars are used for the rehabilitation of concrete structures. These mortars can be applied by hand as well as by machine, in a wet or even a dry spraying shotcrete process. Different kind of mortars delivered as dry-mix mortars with different characteristics and functions are used as the components for concrete rehabilitation systems:

- Primer and adhesion promoter for the reinforced steel (polymer-modified cementitious slurry or epoxy-based coating materials)
- Adhesion-promoter slurry (primer or key-coat) for the concrete to be repaired (polymer-modified cement based slurry)
- Restoration and reprofiling mortar (polymer-modified cement-based mortar)
- Fine stopper or smoothing mortar (polymer-modified cement-based mortar containing fine aggregates)
- Protective and finishing coat (dispersion paints, crack overbridging paints, cementitious waterproofing sealing slurries, etc.).

So far mainly national guidelines implemented by various organizations and regulations are relevant to the requirements for concrete-repair systems [27]. Concrete rehabilitation sys-

tems are subjected to an official approval according to relevant laws and regular quality supervision from independent test institutes.

7. Market Aspects

Precise data on worldwide production and consumption of dry mortars are not available, although for Germany, France, and Italy calculations of the production are known. According to ref. [28] the production of dry mortars in 2001 was 10×10^6 in Germany, 3×10^6 in Italy, and 2.7×10^6 in France, and total consumption in

Western Europe was 30×10^6 [28] or $35-40 \times 10^6$ according to other estimates. Worldwide consumption lies in the range of $50-60 \times 10^6$ [28], or $(60-70) \times 10^6$ according to other estimates.

With increasing local working costs the ratio between dry mortars and mortars mixed on site is increasing, as the application of dry mortars requires less time for the same work. Worldwide dry mortar consumption is increasing. While growth was strong in central Europe from the 1960s to the 1980s, nowadays growth is mainly in Eastern and Southern Europe, as well as in parts of the Asian and Latin American market.

8. References

Specific References

1. A. Konietzko: "The Application of Modern Dry, Factory Mixed, Mortar Products", *ZKG (Zement,Kalk,Gips) International,* **48** (1985) no. 12, 625 – 659.
2. U. Dilger: "Ready-Mixed Mortar Production Plants", *ZKG (Zement,Kalk,Gips) International,* **38** (1985) no. 1, 2 – 6.
3. M. Guldner: "Production and Processing of Dry, Factory-Mixed, Mortars", *ZKG (Zement,Kalk,Gips) International,* **52** (1999) no. 11, 628 – 631.
4. Data from EMO (European Mortar Organisation) and Wacker Polymer Systems GmbH & Co. KG.
5. DIN 1164, Zement mit besonderen Eigenschaften, Zusammensetzung, Anforderungen, Übereinstimmungsnachweis (November 2000).
6. K. Krenkler, *Chemie des Bauwesens, Band 1: Anorganische Chemie,* Springer, Berlin, 1980.
7. J. A. H. Oates, *Lime and Limestone,* Wiley-VCH, Weinheim, 1998.
8. EN 459: Building Lime, Part 1: Definitions, Specifications and Conformity Criteria draft DIN EN 459-1, January 1999; Part 2: Test Methods draft DIN EN 459-2, January 1999; part 3: Conformity Evaluation, draft DIN EN 459-3, January 1999.
9. D. Distler, *Wässrige Polymerdispersionen* Wiley-VCH, Weinheim, 1999.
10. J. Schulze, *Tonindustrie-Zeitung,* **109** (1985) 698.
11. K. Adler, *Schweizer Baublatt* **31** (1988) 44.
12. L. Brandt, *Cellulose Ethers, Ullmann's Encyclopedia of Industrial Chemistry,* 5th ed., **vol. A 5,** (1986) 461 – 488.
13. DIN 18555: Testing of Mortars with Mineral Binders: Part 1: General, Sampling, Test Mortar (September 1982); Part 2: Fresh Mortars Containing Aggregates with Compact Structures, Determination of Consistency, Density and Air Content (September 1982); Part 3: Hardened Mortars; Determination of Binding Strength, Compressive Strength and Density (September 1982); Part 4: Hardened Mortars; Determination of Linear and Transverse Expansion and of Deformation Characteristics of Mortars for Bricklaying by Means of a Static Pressure Test (March 1986); Part 5: Hardened Mortars; Determination of the Bond Shear Strength of Mortars for Bricklaying (March 1986); Part 6: Hardened Mortars; Determination of the Adhesive Tensile Strength (November 1987); Part 7: Fresh Mortars; Determination of the Water Retention Value by the Filter Plate Method (November 1987); Part 8: Fresh Mortars; Determination of the Workability Time and Correction Time of Thin Bed Mortars for Masonry (November 1987); Part 9: Hardened Mortars; Determination of the Mortar Compressive Strength (September 1999).
14. DIN 52617: Determination of the Water Absorption Coefficient of Building Materials (May 1987).
15. Die Wirkung von Fließmitteln in hydraulisch gebundenen Baustoffen. April 1997, SKW-Trostberg.
16. DIN 18557: Factory Mortar – Production, Control and Delivery (November 1997).
17. DIN 1168, Part 1: Gypsum Plasters, Terms and Definition, Types and Application, Delivery and Marking (January 1986); Part 2: Gypsum Building Plasters; Requirements, Testing, Control (July 1975).

18. Draft DIN EN 13279; Gypsum and Gypsum Based Building Plaster; Part 1: Definitions and Requirements (July 1998); Part 2: Test Methods (July 1998).
19. EN 1346: Adhesives for tiles – Determination of open time (March 1999).
20. EN 1308: Adhesives for tiles – Determination of slip (March 1999).
21. EN 1348: Adhesives for Tiles – Determination of Tensile Adhesion Strength for Cementitious Adhesives (March 1999).
22. DIN 18550: Plaster; part 1: Terminology and Requirements (January 1985); Part 2: Plasters and Renderings out of Mortar with Mineral Binders; Execution (January 1985); Part 3: Rendering; Rendering Systems for Thermal Insulation Purposes Made of Mortar Consisting of Mineral Binders and Expanded Polystyrene (EPS) as Aggregate (March 1991); Part 4: Plaster and Rendering: Lightweight Plasters and Renderings; Execution (August 1993).
23. EN 87: Ceramic floor and wall tiles – Definitions, Classification and Marking.
24. Fachverband Wärmedämmverbundsysteme e.V.; Wiesbaden.
25. Wärmeschutzverordnung (WschVO) vom 1.1.1995 der Bundesregierung von Deutschland; Energiesparverordnung (EnEV) 2001.
26. Deutsche Bauchemie e.V.: Instructions on Structural Waterproofing with Cementitious Rigid and Flexible Waterproofing Slurries, July 1993.
27. Bundesanstalt für Straßenwesen, Deutschland: ZTV-SIB, Zusätzliche Technische Vertragsbedingungen und Richtlinien für Schutz und Instandsetzung von Betonbauteilen".
28. Estimation of the Production of Dry Mortars in 2001, European Mortar Organisation, Duisburg, Germany.

"Driers" → Metallic Soaps
Drug Delivery Systems → Pharmaceutical Dosage Forms
Dry Ice → Carbon Dioxide

Pigging Technology

Gerhard Hiltscher, Neustadt a. d. W., Germany
Wolfgang Mühlthaler, Hemsbach a. d. B., Germany

1. **Introduction** 1
2. **Definitions** 2
3. **Pigs** 3
4. **Valves** 4
5. **Piping** 7
6. **Control System** 7
7. **Decision Criteria for Pigging** 8
8. **Example of an Industrial Pigging System** 8
9. **Special Fields of Pigging Technology** . 9
10. **Safety** 11
11. **Producers** 11
12. **References** 11

Pigs are devices which are inserted into and travel through a pipeline, driven by a liquid or gaseous propellant. Installation and operation of pipelines would not be possible without pigging technology. Applications of pigging include sweeping liquids from pipelines, removing condensate, separation of products, inspection, cleaning, and repair. The action of products and/or propellants on pig materials can lead to physical and chemical reactions.Depending upon function, different designs of mechanical and intelligent pipeline pigs are used. Smart pigs are used above all for in-line nondestructive testing of pipelines. Some basic risks must be taken into account when a pigging system is planned.Examples for industrial pigging units are found in food and beverage industry, cosmetics, pharmaceuticals, and in the chemical industry.

1. Introduction

In *pigging*, the contents of a pipeline are pushed by a snug-fitting plug (*pig*) with the goal of removing the product almost completely from the pipeline. The pig is propelled through the pipe by a gas or a liquid (propellant). Apart from the cleaning, pigs can be used to separate different products that are transported in the same pipeline.

Pigging technology was originally developed for the oil industry and has been employed for cleaning pipelines for more than 100 years. Since then, pigging technology has found increasing application in the chemical and other industries.

A *pigging unit* usually consists of the following components:

- The pig
- Piggable pipes with piggable valves
- Pig loading and unloading station
- Propellant supply
- Control system

In the simplest case the pigging unit (Fig. 1) consists of a single pipeline that is traveled by a pig. All components of a pigging unit must be piggable.

Pigs are available in numerous designs, sizes, and materials: From simple spherical pigs, mandrel pigs, and separating pigs to in-line testing and inspection pigs; and from fluid-driven pigs to self-driven camera vehicles. The range of applications of pigs is thus very large.

The most important applications of pigging are:

- Sweeping liquids from pipelines.
- Removing incrustations and deposits.
- Removing condensate (gas pipelines).
- Filling/emptying of a pipeline with plug flow.
- Separation of products, pumped one after the other in the same pipeline (e.g., product A – pig 1 – product B – pig 2 – propellant). This process is called batch pigging.
- Inspection, detection, and observation.
- Cleaning.

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.f20_f01.pub2

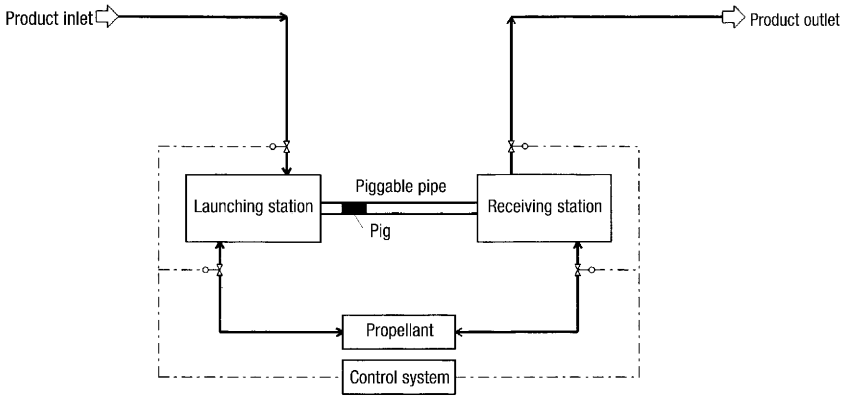

Figure 1. Schematic of a pigging unit

– Measurement and control.
– Repair.

The applications of industrial pigging units encompass four major tasks:

– Several products are pumped through a single pipe
– Product is removed from a pipe
– Rinsing the section of a pipeline between two pigs with a detergent and/or a solvent (tandem pigging)
– Prevention or reduction of foaming

2. Definitions

A *pigging unit* is the total mechanical, and electrical equipment and instrumentation, which is needed for the execution of pig travel. It is a part of an entire plant that cleans, separates, and/or removes liquid substance from at least one pipe. It consists of one or more piggable pipelines including piggable valves, the receiving and launching stations, the control system, and the pig.

Piggable lines can be simple (one piggable line without any branching), branched (consisting several T-piece valves for product inlet and outlet), or with switches (valves with junctions for both pig and product flow).

The *propellant* is the medium which propels the pig. Propellants can be gaseous (compressed air, nitrogen) or liquids such as water (also with additives, inhibitors, or fully demineralized), product, cleaning agents, or solvents.

Pigs, driven by a gas, behave differently to liquid-propelled pigs. Due to their compressibility gaseous driving media lead to uneven pig movement. With liquids as driving medium, ideal plug flow develops behind the pig, and this leads to homogeneous pig movement.

The term *pigging system* refers to the different pigging procedures that are possible in a pigging unit, i.e., the temporal sequence of individual operating steps. A distribution is made between open and closed pigging systems, and between one- and two-pig systems.

In an open pigging system (removable pig) the pig can travel through the pipe only in one direction (Fig. 2). At the receiving station the pig is removed and returned externally to its starting station.

In open pigging systems, pigs with conical, cup-shaped seals are generally used which can be driven only in one direction. Often several pigs are present at the launching station, which are collected at the receiving station for return. The cleaning of the pigs is carried out manually outside of the pigging unit. Open systems are

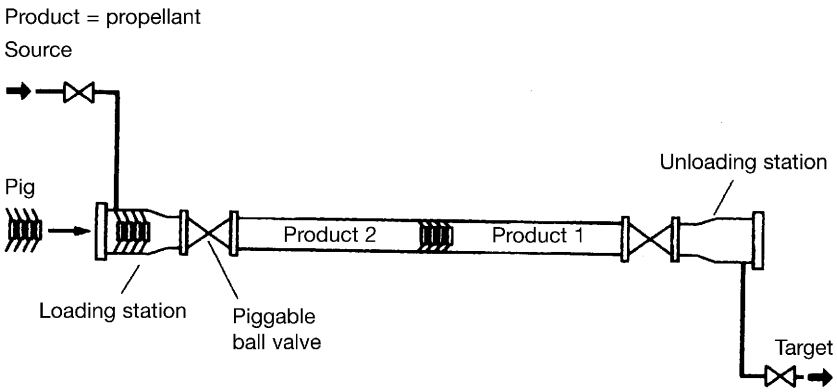

Figure 2. Open pigging system (schematic)

particularly suitable for long lines (> 1 km) in the chemical industry, e.g., from a tank farm to a ship at a loading jetty or long-distance pipelines.

In a closed pigging system the pig remains during its total service life in the line. Only pigs whose form enables movement in both directions are suitable. The closed pigging system is versatile, e.g., piggable diverter valves can be used to construct a branched pigging unit.

One-pig systems (OPS, Fig. 3 A) can be open or closed. In two-pig systems (TPS, Fig. 3 B) the two pigs can be driven in the same direction with or without interpig spacing or in opposite directions through the line.

3. Pigs

Pigs are devices which are inserted into and travel through a pipeline, driven by a liquid or gaseous propellant. The pigs slide on a thin liquid film (micrometer range) and clean the pipe with two or more narrow lips. The pigs are thicker than the internal diameter of the pipe and are pressed into the piping. The resulting strain which depends on the diameter and type of pig and amounts to ca. 3 % prevents aquaplaning. Thus a high degree of cleaning is ensured.

The chemical and physical characteristics of the products to be pigged influence the selection of the *pig materials*. Depending on the application different pig materials are used. The most frequently employed materials for pigs are elastomeric polymers.

Pig designs can be classified as one-piece and multicomponent. In one-piece pigs, body and cleaning lips from a unit, while multicomponent pigs consist of a central body that can be equipped with different cleaning elements.

Spherical pigs, solid cast pigs, lipped pigs, cylindrical pigs, pigs with exchangeable lips, conical seal pigs, and various specialty pigs are available. Depending on the required degree of cleaning the size of the pipe, and the system pressure, a suitable pig is selected.

The *solid cast pig*, the main features of which are described in Figure 4, is among the most frequently used one-piece pigs. It is adequate for most pigging processes. Solid cast pigs are very durable and are equipped with two inflexible lips that clean of the pipe. Solid cast pigs are pressed into the pipe with a prestress and can be driven bidirectionally. High service lifetimes can be achieved. Polyurethane has the best mechanical properties for these pigs.

Exchangeable-lip pigs (Fig. 5) are the simplest examples of multicomponent pigs. They have a solid body, made of plastic or metal, with two or more removable flexible lips. Af-

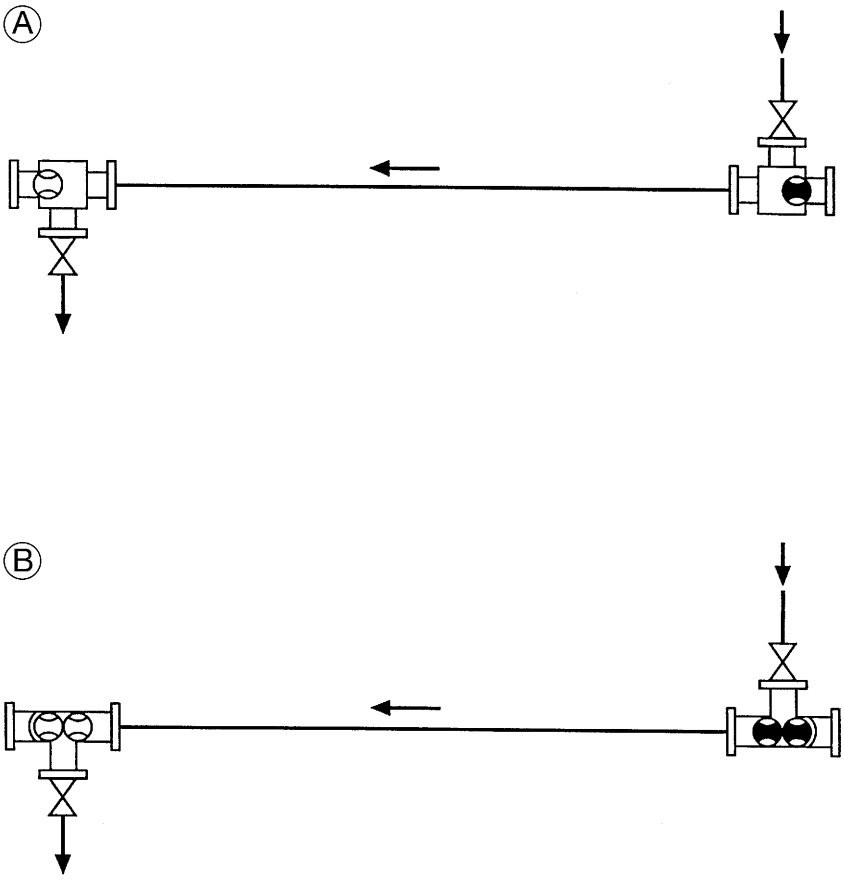

Figure 3. A) Closed pigging system (OPS); B) Closed pigging system (TPS)

ter changing the worn lips, the solid body can be used again.

Special pigs are used in the oil industry, in sterile areas, in the foodstuffs industry, and with bulk materials for cleaning purposes and for blockage-free transport.

The different chemical and physical characteristics of piggable products crucially affect the selection of materials for pigs.

The action of products and/or propellants on pig materials can lead to physical and chemical reactions. The results of physical processes are revealed by changes in weight, volume, and dimensions, and the influence of chemical reactions by changes in hardness, ultimate tensile strength, and fracture strain. If no suitable pig material is found, then an industrial pigging unit must be dispensed with.

4. Valves

In a piggable piping system the selection of the piggable valves is particularly important. They must be optimally adapted to the application. Apart from the general requirements for valves, such as tightness, low leakage, smoothness of operation, and precision the piggability of the valves must be ensured, e.g., by having:

- The same internal diameter as the pipeline
- A centerable flange
- Accurate adjustability of plugs in stopcocks and switches
- Guiding rods at valve outlets

The piggable valves are classified as standard and special valves:

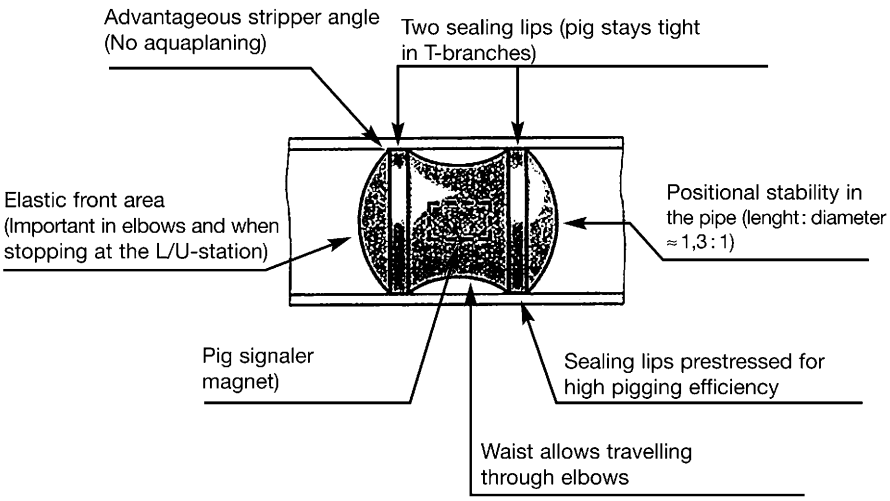

Figure 4. Solid cast pig (company I.S.T.)

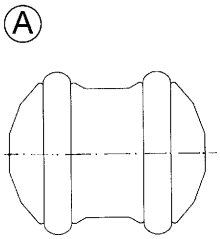

Figure 5. A) Exchangeable-lip pig (company Pfeiffer); B) Exchangeable-lip pig (company Kiesel)

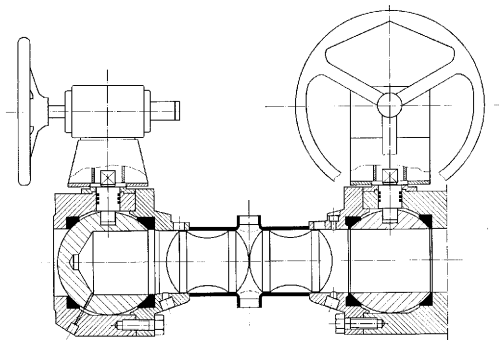

Figure 6. Loading and unloading station (company Pfeiffer)

Standard valve	Special valves
– Stations	– Crossing of two piggable pipes
– Branches	– Pig receiving station for loading arms
– Pig traps	– Pig receiving station for filling valves
– Switches	– Manifold
	– Valves for the sterile areas

The *pig loading and unloading station* (Fig. 6) can be at any end of the pigging system. At the loading and unloading station the pig is introduced into or removed from the pigging system, and can also be driven through the pigging system by a propellant.

Before the pig is changed, the valve or the pipe must be completely depressurized or separated from the piping system by a safety valve. Insertion and removal of pigs must be easy, fast, and safe. Fittings for cleaning agents, as well as pig sensors or mechanical pig tracers for detecting the pig, must be incorporated in the valve.

The station simplifies the pig loading and unloading substantially by means of the integrated shut-off ball valve. The ball has a blind hole and can be rotated by 180 ° for removing and inserting the pig. The diameter of the blind hole is larger than the outer diameter of the pig and the

internal diameter of the piggable pipe so that the pig can be inserted easily. The transition between valve and pipe is conical, so that the pressure of the propellant pushes the pig into the piggable pipe.

If product is to be removed from or introduced into a piggable pipe, then a *branch* is required. Usually the branch is arranged vertically with respect to the pipe axis. Examples of branches are the T-ring valve and the throughput piston valve.

The *T-ring valve* (Fig. 7) is a dead-space-free special valve that functions as a check valve, piggable T-piece, and pig trap. The straight-through passage can be pigged without a dead space, whereby the valve is closed. The T-branch can be shut off by the sliding ring, which also stops the pig. A further check valve at the T-piece is not required.

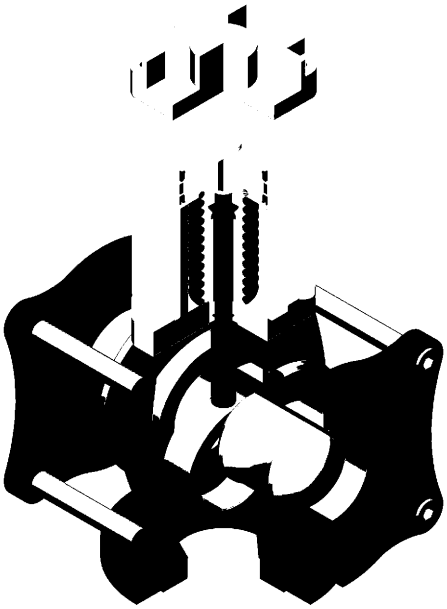

Figure 7. T-ring valve (company I.S.T.)

The *throughput piston valve* (Fig. 8) combines the functions of a piston valve and a pig retainer or pig stopper in one valve. In the closed state the piggable pipe runs smoothly without dead space through the piston valve. In the open state the product can be withdrawn downwards and the pig thereby is held on the piston of the valve.

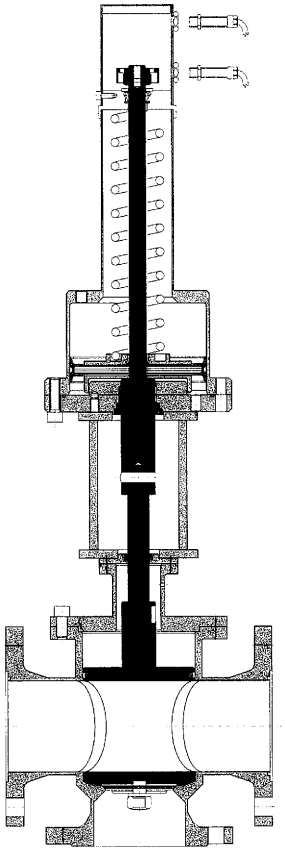

Figure 8. Throughput piston valve (company Kiesel)

In contrast to product branches, *switches* divert both product and pig. An example is the three-way switch (Fig. 9).

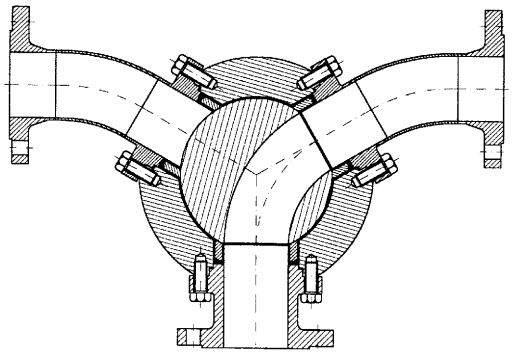

Figure 9. Three-way switch (company Pfeiffer)

Manifolds allow piggable pipes to be connected with nonpiggable pipes. Manifolds are required in multiproduct units for mixing, filling, and shipping procedures.

The full-system manifold (Fig. 10) is closed and piggable on one side. The full-system section consists of nonpiggable rectangular ducts for product feeding. At right angles to these ducts are the piggable pipes, which are connected at the intersection by means of ring valve. This ring valve opens the connection from the product feed duct to the piggable part of the manifold and simultaneously acts as a pig receiving station. Manifolds make it possible to connect and distribute up to 50 feed pipes to as many as 20 piggable pipes in a single module.

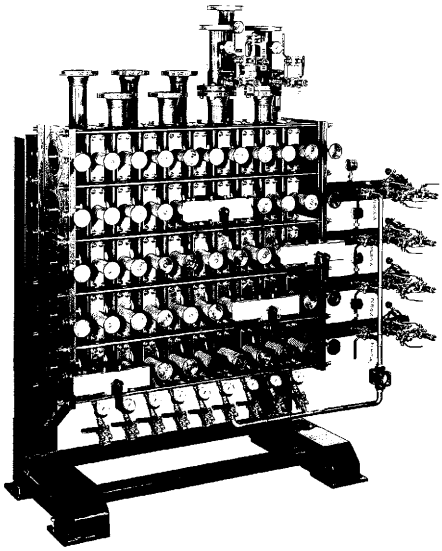

Figure 10. Full-system manifold (company I.S.T.)

5. Piping

In a pigging system the piggable pipes, pipe bends, and joints are of major importance. The quality of the overall system determines the cleanness of the piping after a pig travel.

Therefore, all important boundary conditions are specified in standard DIN 2430 (available in English): Piping for Pigging Systems: Part 1: Pipes and Tube Bends, Part 2: Fittings, Part 3: Testing Prior to Commissioning.

The preferred materials (DIN 2430) are the stainless steels 1.4307, 1.4404, as well as AISI 304L, and AISI 316L.

The characteristics specified by the manufacturer are checked in the important test pigging.

In the test pigging, propellant pressure and time are measured, and thus pressure peaks, caused, e.g., by flange misalignment, welding seam dips, pipe bends, valves, dents, etc., can be detected and located. The ratio of back pressure to peak pressure is a measure for the quality of manufacturing and erection of the piggable pipe.

Test pigging is a component of the function guarantee and is required by all companies which design, supply, and erect industrial pigging units.

6. Control System

Planning the control system for a pigging unit requires knowledge of this special technique. Control of a pigging unit should preferentially be automatic by means of a distributed control system (DCS) or, for smaller systems, a programmable logic controller (PLC). A field bus system connects the field devices (actuators and sensors) directly to the control system. Monitoring and operating components are necessary to supervise the locations and motions of the pigs and to check the valve settings, pump running information, and tank level indicators.

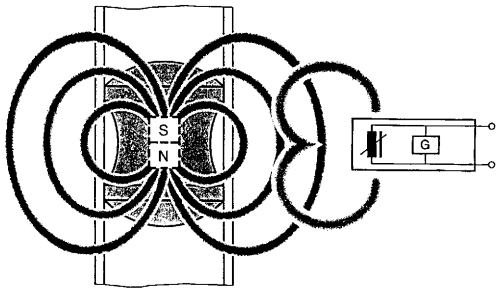

Figure 11. Principle of a magnetic pig indicator

One of the most important sensors of a pigging system is the pig detector (Fig. 11). For this purpose a permanent magnet is installed inside

the pig. Outside the line magnetic-inductive indicators are mounted at suitable locations (e.g., pig stations, junctions). To avoid faulty measurements, the material for lines and valves must not be magnetizable (paramagnetic).

All the other sensors (pressure, level, temperature) found in a pigging system are regular commercial instruments. Recent developments in the instrument technique of the Coriolis mass flow meters permits this technology to be used directly in piggable lines. Ultrasonic flow measurements, which are applied outside from the pipe, can also be used.

7. Decision Criteria for Pigging

Apart from technical necessity, the economy of a pigging system is the focus of attention. It is above all crucial, that the new technology brings an advantage on a long-term basis over a competitor. Therefore, all operational criteria must be checked.

In the case of long pipes, multiproduct systems, or batch operation the economic advantages of a pigging system become particularly apparent:

- A single line is used for several products (saves on investment costs and space requirements)
- Easy emptying of the line with products which can freeze, condense, decompose, or polymerize
- Dispensing with insulation and/or tracing (line heating)
- Time savings compared to emptying the line by a manual procedure
- No rinsing procedures and/or substantially lower amounts of cleansing agents, lower costs for wastewater treatment, lower incineration costs, reduction of product losses
- No need for slopes and syphon—free installation to be able to empty the pipeline completely

8. Example of an Industrial Pigging System

A chemical production plant produces several 100 000 t/a aqueous urea – formaldehyde condensation products, mainly for the chipboard industry. The entire complex (Fig. 12) consists of:

- The production plant consisting of production intermediate vessels and a manifold station inside the building.
- Two large tank farms, connected by pipes to the process building. The pipes are up to 300 m long.
- Loading stations for tank trucks and rail cars to be filled from the tank farm.

The individual pigging lines were equipped with the two-pig systems without any detergents. In the production plant, on the pipe racks, and inside the tank farms there are ca. 2000 m of piggable lines. Within the production plant there are the following 4-inch piggable lines:

- From vessel group A to the manifold (ten three-way switches)
- From vessel group B to the manifold
- From the manifold to tank farms I and II

Between the tank farms and the loading stations there are the following 6-inch piggable lines:

- Between tank farms I and II
- From three groups of tanks to tank-truck loading
- From three groups of tanks to rail-car loading

Other conditions:

- Propellant: product or compressed air (0.4 MPa g) (g means gauge, i.e., overpressure)
- Pig-travel frequency: several times daily
- Cleaning degree: D [1]
- Solid cast pigs
- Two-pig system

The pigging systems has been in successful operation for more than 20 years, and has completely fulfilled its purpose:

- Using one pipe for several products of a product family
- Emptying and cleaning of the pipes by almost complete removal of the products
- Rapid product change and perfect separation of the products among themselves
- Avoidance of product loss by rinsing procedures
- Prevention of condensate developing and thus of blocking of pipes
- Substantial savings in wastewater costs
- Maintenance of high product quality

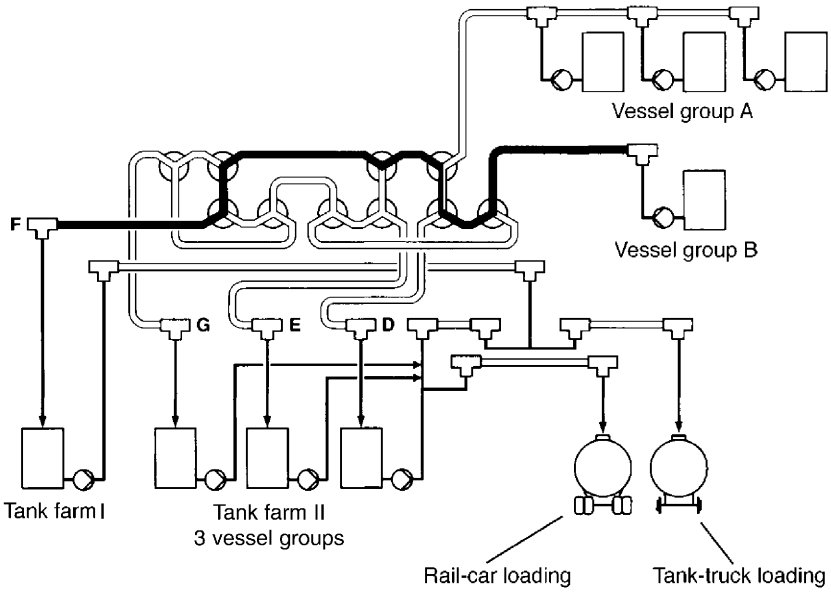

Figure 12. Example of a branched pigging system

9. Special Fields of Pigging Technology

Pigging in Hygienic and Sterile Applications. The term "hygienic and sterile pigging" refers to all applications in which special requirements must be fullfilled concerning cleanness and hygiene up to the extremely high requirements (e.g., GMP, FDA regulations) of sterile technology.

This particularly applies to the food and beverage, cosmetics, and pharmaceutical industries, including biochemistry and genetic engineering, as well as to special applications in chemical plants.

In fully automated processes in the food and beverage industry, the cleaning procedure is of the same importance as the production itself for hygienic reasons. Cleaning must take place regularly, often daily, to guarantee the required freedom from microorganisms. The products are predominantly of high value, so that product losses must be minimized. Cyclic cleaning procedures lead to high disposal costs.

Pipes in sterile technology must be extremely smooth inside and outside; the used fittings and valves have to be free of any dead space. Pigging, especially with closed piping systems, is especially suitable as an automable cleaning method for pipes in sterile plants.

Examples applications for industrial pigging units:

Food industry:	Yogurt, cheese, mustard, paste, jam, chocolate
Beverage industry:	Dairies, breweries, and wine cellars, syrups, fruit juices and cola concentrates
Cosmetics:	Sun screens, shampoos, lotions, creams, toothpastes
Pharmaceuticals:	Ointments, cough mixtures
Chemical industry:	Cleaning agents, detergents, vitamins, dispersions, paints and coatings

Pipeline pigging is the oldest field of pig applications.

Pipelines are used for the transport of primary energy and pumpable bulk goods. They can be used exclusively for the transport of only one product, but there are also multiproduct pipelines (e.g., different crude oils).

Pipelines (long distance lines) have other characteristics than the pipes in chemical plants. Apart from the length and the diameter these are:

– Pipelines are laid in the landscape and thus they are not free of syphons (i.e.,

not automatically emptiable). An installation without syphons is absolutely impossible, while within the chemical industry a limited syphon-free pipe installation can be achieved. The volumes of the syphons are very large (field-erected bends).
- The holdup of a pipeline is extremely large. Mass and/or flow meters are prescribed for balancing.
- Large parts of the pipeline are not accessible from the outside; this particularly applies to the offshore region.
- Industrial pipes contain substantially more fittings (bends, branches) and valves.
- Pipelines are used for few products: crude oil, refined oil, natural gas, and water, whereas the pipes in industrial plants are used for the total spectrum of chemicals.
- Pigging technology and pigging systems are substantially more complex with industrial pigging units.

A clear distinction between long-distance lines and the pipes of a chemical plant is not possible, since in practice a wide overlap exists. Also within chemical plants there can be extremely long pipelines of large nominal size, for example, naphtha lines for the cracker feed between a ship unloading facility and the process plant.

Pipelines are divided into onshore, offshore, and subsea lines, each having special requirements for strength, resistance, and installation technology. For the operation and inspection of such lines the pig must fulfill the most diverse functions.

Usually a pipeline is divided into several sections, separated by stations. In pipeline technology these stations have the ability shut the line off section by section, to create the line pressure (compressor and/or pumping stations), as well as control measurements. The station also permits pig travel. The longest distance between two pigging stations at present is approx. 800 km.

Installation and operation of pipelines would not be possible without pigging technology. Numerous tasks can only be solved economically with pigs; pipeline technology is closely connected with pigging technology, and every pipeline is equipped with pig stations. Therefore, the first applications of pigs coincided with the building of the first pipelines. Around 1870 in the USA the first long-distance pipelines for crude oil were laid, and the term pipeline pig was used for the first time.

Pigs are used in the erection phase and later when operating, inspecting, maintaining, and repairing the pipeline:

- During fabrication: Removal of coarse dirt, removal of liquids, drying, final inspection
- During start up: Complete flooding for pressure testing, calibration of volumetric flow meters
- During operation: Cleaning (removal of deposits), condensate removal, separating different products
- During inspection: Checking geometry; detection of corrosion, erosion, and cracks; leakage detection
- During repair: In situ in-line coating, corrosion inhibition, plugging of pipeline sections, shutting off lines

Depending upon function, different designs of mechanical and intelligent *pipeline pigs* are used. Separation pigs and cleaning pigs belong to the group of mechanical pigs, whereas detection pigs and inspection pigs are classified as intelligent pigs.

Different methods are available for pig detection in pipelines: mechanical, ultrasonic, and radioactive detection, and transmitter/receiver systems. Since pipelines mainly consist of ferritic steel, pig detection by means of built-in permanent magnets, as in industrial pigging units, is not possible.

Solid cast pigs for use in pipelines are mostly made of open-cellul polyurethane, rather than the solid plastics of the industrial pigs. Polyurethane foam pigs are available in different densities and with various configurations of solid polyurethane strips bonded to the surface. They generally have a bulletlike shape and are often used for simple cleaning functions (wiping, drying) as unindirectional pigs. However, heavy-duty cleaning procedures, such as the removal of wax or other firmly adhering deposits can only be achieved by pigs with metal brushes.

Over to their size, pigs for pipelines are often multicomponent pigs (mandrel pigs): a central body system or mandrel can be equipped with various components (e.g., different disks, seal-

ing rims and cups, scrapers, and brushes). Worn parts can be exchanged in this way.

Pigs which do not consist solely of mechanical components (mechanical pigs), but also have some electrical/electronic parts are called intelligent (or smart) pigs. They are able to measure, to process, to store, or to transmit data. Smart pigs are used above all for in-line nondestructive testing of pipelines (e.g., magnetic flux leakage, ultrasound, and eddy current testing). With smart pigs, crack detection, leak detection, measurement and location of pipe defects, detection of corrosion and metal loss, and geometry measurement are possible. Photographic-inspection and video-monitoring systems are also available.

10. Safety

The risks that occur in pigging systems can be assigned primarily to three causes:

1) The kinetic energy of the traveling pig
2) The pressure and transient pulses caused by the propellant
3) Formation of explosive air – vapor mixtures when inflammable liquids are pigged with air.

These basic risks must be taken into account when a pigging system is planned. For further information, see [1] and [2].

11. Producers

Some major suppliers of pigging equipment in Germany are listed in the following:

– I.S.T. Pigging Technology GmbH, Schierenbergstr. 74, D-22145 Hamburg
– G. A. Kiesel GmbH, Wannenäckerstr. 20, D-74078 Heilbronn
– Pfeiffer Chemie-Armaturenbau GmbH, Hooghe Weg 41, D-47906 Kempen

12. References

1. G. Hiltscher, W. Mühlthaler, J. Smits, *Molchtechnik: Grundlagen, Komponenten, Anwendungstechnik*, Wiley-VCH, Weinheim 1999.
2. G. Hiltscher, W. Mühlthaler, J. Smits, *Industrial Pigging Technology, Fundamentals, Components, Applications*, Wiley-VCH, Weinheim 2003.

Superabsorbents

Markus Frank, Stockhausen GmbH & Co KG, Krefeld, Germany

1. Introduction 1
2. Definition 1
3. Theory of Polyelectrolyte Networks 2
3.1. Elasticity 2
3.2. Swelling 3
3.3. Phase Transition 4
4. Chemistry and Production of Superabsorbents 5
4.1. Raw Materials 5
4.2. Radical Solution Polymerization . . 6
4.2.1. Principles 6
4.2.2. Production Processes 7
4.3. Inverse Suspension Polymerization 7
4.4. Surface Treatment 8
5. Properties, Characterization, and Analysis 8
5.1. Absorption Capacity 9
5.2. Shear Modulus 10
5.3. Kinetics of Absorption 11
5.4. Swelling under Load 11
5.5. Permeability 12
5.6. Test Methods 13
5.7. Stability of Superabsorbents 13
6. Uses 13
6.1. Personal Hygiene Products 13
6.1.1. Diapers 14
6.1.2. Feminine Hygiene Products 14
6.2. Packaging Applications 15
6.3. Soil Conditioners 15
6.4. Cable Applications 15
6.5. Civil Engineering and Construction 16
6.6. Other Applications 16
7. Other Forms of Superabsorbent Polymers 16
7.1. Fibers 17
7.2. Foams and Films 17
8. Environmental Aspects and Toxicology 17
9. Economic Aspects 18
10. References 18

1. Introduction

Superabsorbents are materials that can absorb up to 1000 times their mass of water. They are usually white, crystalline powders. When they come into contact with aqueous liquids, within a short period of time the powder changes into a gel-like substance and the liquid is absorbed. In this context the absorbency, however, depends very strongly on the type of aqueous liquid. In water dissolved ions in particular do strongly reduce the absorbency. Absorbents for non aqueous liquids are not part of the concept superabsorbents. The interested reader is referred to the relevant literature [1–4]

The development and commercial success of superabsorbents is closely connected with disposable diapers. Although the first superabsorbents were developed as water reservoirs for soils and as carriers for active ingredients, it was their use in diapers in the 1980s that helped in their breakthrough.

The first superabsorbents were polyacrylonitriles, grafted on to starch and subsequently saponified. Later, superabsorbents based on acrylic acid quickly gained recognition for economic reasons and due to their superior properties. A detailed overview on superabsorbents can be found in [5,6].

2. Definition

In the literature superabsorbents appear under different names. Hydrogel, polyelectrolyte gel, water-swellable polymer, water-absorbent polymer, superabsorbent material (SAM), and superabsorbent polymer (SAP) are the most frequent designations. Gross used the designation xerogellant, which describes both a dry gel and the behavior of the material [7].

Superabsorbents are capable of absorbing large quantities of aqueous fluids spontaneously and rapidly. The aqueous fluid is strongly retained and is not released mechanically. While swelling they essentially keep their original shape (Fig. 1). They only change in their dimensions and rheological behavior: a brittle solid material becomes a gel.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.f25_f01

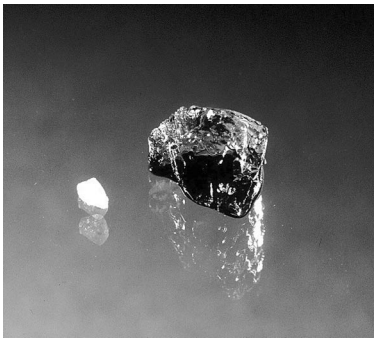

Figure 1. Superabsorbent particle after production and fully swollen in 0.9 % NaCl solution

Absorbency and Superabsorbency Absorption is a widespread common phenomenon. Besides surface adsorption there are two principal mechanisms of absorbence. One is the physical filling up of the open space of a porous media with capillary pressure as driving force. The other is diffusion of the liquid into the absorbent material itself. In this case the driving force is the difference in chemical potential inside and outside of the absorbent material.

The porous frame work of conventional absorbent materials, such as foams, absorbent papers, and textiles, can be formed by fibers, inorganic substances, or polymers. These materials can take up only a small amount of liquid. Liquids wet the surface and are transported to the interior by pores and channels. The extent and rate of absorption depend on the wettability, the pore volume, the length and radius of the pores, the viscosity and surface tension, as well as external forces such as gravitation and pressure differences. The transport of liquid through a porous material is commonly described by Darcy's law. It is valid for a linear and slow steady flow.

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{kA\Delta P}{\eta L_0} \tag{1}$$

The volumetric flow rate $\mathrm{d}v/\mathrm{d}t$ is given in terms of the cross-sectional area A of the sample, the length L_0 of the sample in the direction of flow, the permeability k of the medium, the viscosity η of the fluid and the pressure difference ΔP driving the flow.

Before superabsorbents were used, the absorbent cores of diapers consisted of only cellulose fluff or layers of absorbent papers. The main difference of these materials from superabsorbents is that they release most of the fluid under mechanical pressure. An extensive review of absorbent materials can be found in [8].

The absorbency of superabsorbents is chemical in nature. When an aqueous liquid comes into contact with a superabsorbent, it is spontaneously absorbed until the difference in chemical potential inside and outside the superabsorbent is balanced. The absorbed liquid can hardly be released mechanically. Water can only be removed from common superabsorbents by drying, applying high pressures, or deteriorating the solvent quality of water by adding water soluble salts or water miscible organic liquids.

3. Theory of Polyelectrolyte Networks

The majority of superabsorbents produced today are partially neutralized polyacrylates. Therefore, the theory of superabsorbents must deal with polyelectrolyte networks. They have long been studied, but a precise mathematical description is difficult and is still incomplete. The discussion of all the theories is beyond the scope of this article; for further information, see [9–12].

The most important commercial evaluation criteria of superabsorbents are the amount of aqueous liquid that can be absorbed, the rate at which it is absorbed, and the absorption of aqueous liquids under compression or under an external load. The theories try to relate the shear modulus G and the volumetric swelling ratio q to the structure of the polymer network. The structural parameters of a polyelectrolyte network are the number of cross-links per unit volume ν^*, the mean number of statistical segments between two junction points N, the interaction parameter between solvent and polymer chain χ, the molecular mass of the polymer chain per free counterion M_2, and the number of charges per polyion chain and their valences.

3.1. Elasticity

Superabsorbents are polyelectrolyte networks. Like other polymer networks they are rubber-elastic materials. Therefore, they are mostly described by models of rubber elasticity with additional ionic contributions. Modern theories use

the scaling concept for the description of polyelectrolyte networks.

The classical theories of rubber elasticity are based on the assumption that the total change in the free energy of the network's is composed cumulatively of the changes in free energies of the individual chains. There are two fundamental models, the model of the affine network [13] and the model of the phantom network [14]. In both models the polymer chains have no residual volume and therefore can interpenetrate unhindered. Whereas in the affine network model the fluctuation of the cross-links is suppressed completely due to the presence of the neighboring chains, the cross-links of the phantom network model can fluctuate around their mean position due to Brownian motion. Further developments of the phantom network [15–17] consider the influences of chain loops on the fluctuation of the cross-links. The model is also called the constrained junction model. As superabsorbents are generally cross-linked in solution, the swelling degree q_c, which was present during cross-linking must be considered. The relative swelling degree q_r is the swelling from the state of cross-linkage to the state of equilibrium.

$$q_r = q/q_c \tag{2}$$

The distribution of the chain-end distances is described in the simplest case by a Gaussian error function. In personal-care applications the polymer chains are not greatly extended because the swelling is limited due to the high ionic strength of the urine that is to be absorbed (see Sec. 3.2). Therefore, a Gaussian distribution leads to a good prediction of the dependency of the degree of swelling and the shear modulus.

$$G = A\eta\nu^* \mathrm{R}Tq^{-1/3}q_c^{-2/3} \tag{3}$$

R = general gas constant; T = temperature; A = structural factor, $A = (1 - 2/f)$, f = functionality of the cross-linker; and $\eta = \langle r^2\rangle_i/\langle r^2\rangle_0$ is the ratio of the end-to-end distances of cross-linked and free chains. The number of cross-links per unit volume ν^* can be calculated from the ratio of the concentration of cross-linker to monomer if the polymer network is ideal. In the case of superabsorbents, formation of the network by radical polymerization results in network defects. Besides loop formation and chain entanglements, chains are present that are attached to the network by only one link. Chains or small parts of the network which are not linked to the rest of the network can be extracted. The rate of extraction depends on the molecular mass and the degree of branching of the polymers. These extractable polymers do not contribute to the shear modulus but can contribute to swelling if they remain in the network during swelling. In addition the cross-linking density is normally heterogeneous, and the molecular mass of the chains between two cross-links M_c varies.

When polyelectrolyte networks swell strongly, the finite flexibility of the chains must be considered. Following the approach of Oppermann and Schröder [18], based on non-Gaussion chain statistics, the shear modulus G was obtained from the initial slope of the stress – strain equation [19]:

$$G = A\mathrm{R}T\nu^* q_c^{-1} q_r^{-1/3} \frac{1}{6}\left(\frac{L^{-1}(\gamma)}{\gamma} + \frac{\partial L^{-1}(\gamma)}{\partial\gamma}\right) \tag{4}$$

$\gamma = q_r^{1/3}N^{-1/2}$, L^{-1} = inverse Langevin function; $L(\gamma) = \coth\gamma - 1/\gamma$; $\eta = 1$.

The dependency of the shear modulus G on the swelling ratio q is influenced by the the number of elastic effective chains in the network ν^* and the mean number of statistical segments between two junction points N. N and ν^* can be evaluated by fitting the theoretical curve to the experimental data.

3.2. Swelling

When brought into contact with an aqueous fluid, cross-linked polyelectrolyte polymers swell and form a gel. The cross-linkage of the polymer chains hinders the dissolution of the polymer. The widening of the network decreases the degrees of freedom of the polymer chains of the network, that is, the entropy is reduced. With uncharged gels the entropy reduction is compensated by the energy which is generated by mixing polymer and solvent. The major driving force in the swelling of polyelecrolyte gels is the osmotic pressure due to fixed charges along the polymer chains and the free counterions which

are present when an aqueous liquid is added. An equilibrium between the gel phase and the surrounding solvent is reached, and the chemical potentials of all mobile components compensate each other.

The considerations are reduced to the chemical potential of the solvent water μ_1. In general three contributions are considered:

$$\Delta\mu_{1,\mathrm{mix}} + \Delta\mu_{1,\mathrm{elast}} + \Delta\mu_{1,\mathrm{ion}} = 0 \quad (5)$$

$\Delta\mu_{1,\mathrm{mix}}$ describes the mixture of the polymer with the solvents and results, in accordance with the Flory – Huggins theory [20] in the following relationship:

$$\Delta\mu_{1,\mathrm{mix}} = \mathrm{R}T\left[ln\left(1-q^{-1}\right)+q^{-1}+\chi q^{-2}\right] \quad (6)$$

The interaction parameter χ depends on the concentration, the temperature, and the pressure.

Hooper et al. developed a more realistic model for the interactions in polymer – water mixtures that considers hydrogen bonds and dipolar interactions [21]; however, it is more difficult to handle mathematically and contains further parameters which must be determined experimentally.

$\Delta\mu_{1,\mathrm{elast}}$ represents the expansion of the polymer network during the swelling process and can be described by the theories and assumptions on rubber elasticity [22, 23].

$$\Delta\mu_{1,\mathrm{elast}} =$$

$$1/3A\nu^{*}\mathrm{R}TV_{01}q_{\mathrm{c}}^{-1}q_{\mathrm{r}}^{-2/3}N^{1/2}L^{-1}\left(q_{\mathrm{r}}^{1/3}N^{1/2}\right) \quad (7)$$

To maintain electroneutrality, the charges along the polymer chains must be neutralized by the same number of counterions. This can lead to differences between the concentration of the counterions in the gel and in the surrounding solution. The concentration can only be balanced by diffusion of the solvent into the gel. This is the reason for the strong dependence of the swelling on the ionic strength of the surrounding solution. The lower the ionic strength the higher the degree of swelling. This behavior can also be explained using the Donnan equilibrium (→ Ion Exchangers, Chap. 6.1.), in which the phase boundary of the gel to the surrounding solvent is regarded as a kind of semipermeable membrane. If an anionic polyelectrolyte network is in a solution of a monovalent counterion and if the concentration of the ions c_{s} is used instead of the activities, this results in the following expression for the ionic contribution to the chemical potential $\Delta\mu_{1,\mathrm{ion}}$.

$$\Delta\mu_{1,\mathrm{ion}} = 2\mathrm{R}TV_{01}\left[c_{\mathrm{s}} - \sqrt{c_{\mathrm{s}}^2 + \left(\frac{r}{2qM_2}\right)^2}\right] \quad (8)$$

where ρ is the density of the dry polymer.

Further contributions to the description of the swelling behavior of ionic gels which are discussed in literature are the influence of the charges on the chains on the conformation and the stiffness of the chains, as well as additional electrostatic interactions. Since the charges on a polymer chain repel each other, the stiffness of the chain (the number of statistical segments N) is a function of the effective charge density and ionic strength in the gel at equilibrium.

More recent theoretical work describes the elasticity and the swelling behavior of polymer networks with scale arguments [24–26]. Rubinstein et al. [27] applied new discoveries about the relationship of chain configurations and the shear modulus G' on polyelectrolyte gels. These models describe the experimental data in part more accurately than the classical theories.

From the above equations it is difficult to derive an equation that directly relates swelling to the important parameters. Buchholz derived a simple equation assuming a network randomly cross-linked by a four functional cross–linker and neglecting the ionic effects [28].

$$q_r = \left(\frac{2kM_{\mathrm{c}}}{r_0\nu_{20}^{2/3}\left(1-3M_{\mathrm{c}}/M_{\mathrm{n}}\right)}\right)^{3/5} \quad (9)$$

$$k = \frac{(1/2-\chi)}{V_1} \quad (10)$$

ν_{20} = polymer concentration at the time of preparation during cross-linking, V_1 = volume of the swelling liquid, ρ_0 = density of the dry polymer, M_{n} = molecular wheight of the backbone, M_{c} = average molecular weight between junctions.

3.3. Phase Transition

Ionic polymer gels undergo discontinuous volume phase transitions in response to small changes in, for instance, temperature [29], sol-

vent composition [30], or when an external electric field is applied [31]. The collapse of polyelectrolyte gels in aqueous organic solvent mixtures has been studied extensively. Especially the effect of the pH and the ionic strength of the solvent on the phase transition was of interest [32, 33]. Prerequisite for the discontinuous volume phase transition is that the gel has enough positive internal osmotic pressure and that the solvent in which the gel is immersed is sufficiently poor [34]. The phase transition has been ascribed to the formation of ion clusters or ionomerlike conformations driven by the electrostatic attractions among ion pairs consisting of charged polymers and counterions [35].

Tanaka [36–38] used the Flory – Huggins formula and the Donnan equilibrium theory to describe the phase transition of polyelectrolyte networks. The occurrence and sharpness of the phase transition is characterized by the degree of ionization of chains in the gel f and the excess free energy of contact between polymer segments ΔF_s, which is influenced by the amount of organic solvent and added salt. When f exceeds a critical value f_c, the shrinkage of the gel with increasing amount of organic solvent becomes discontinuous.

$$\tau = -\frac{S_0 \emptyset_0}{r^2}\left[(2f+1)\,r - 2r^{1/3}\right] + 1 + \frac{2}{\emptyset_0}\frac{1}{r} + \frac{2ln\,(1-\emptyset_0 r)}{\emptyset_0^2 r^2} \qquad (11)$$

$$S_0 = u\nu/A\emptyset_0^3$$

$\emptyset$, $\emptyset_0$ = initial, final equilibrium volume fractions of the network, r = swelling ratio $\emptyset/\emptyset_0$, A = Avogadro's number, υ = solvent molar volume, ν = number of constituent chains per unit volume at $\emptyset = \emptyset_0$.

4. Chemistry and Production of Superabsorbents

The term superabsorbent is always used in connection with the absorption of aqueous liquids. Hence, modern superabsorbents are hydrophilic polymers. They are categorized into ionic and nonionic polymers, whereby the ionic polymers are further categorized into cationic, anionic, and amphoteric polymers. Acrylic acid is the basis for almost all commercial superabsorbents. They are generally manufactured by the cross-linking radical copolymerization of partly neutralized acrylic acid in water with a cross-linker having two or more polymerizable groups [39]. Linear polyacrylamides and poly(acrylic acids) are discussed in → Polyacrylamides and Poly(Acrylic Acids). The statements made largely also apply to the synthesis of cross-linked polyacrylic acids. For the synthesis of absorbent polymers based on other monomers, see [40–42].

4.1. Raw Materials

Monomers. Good polymer properties can only be attained with high-purity acrylic acid. Dimers of acrylic acid are formed during storage by Michael addition of one monomer molecule across the double bond of another. When the dimer is incorporated into the poly(acrylic acid) network the reaction can be reversed at high temperatures during the drying step, and one molecule of acrylic acid is cleaved off. Acrylic acid is typically inhibited with ca. 200 ppm of the monomethyl ether of hydroquinone (MEHQ). The effectiveness of MEHQ is increased in the presence of oxygen, which is added during storage but has to be removed to facilitate polymerization.

Besides acrylic acid, in principal all hydrophilic polymerizable substances come into consideration as monomers for superabsorbents. In addition, hydrophobic polymerizable substances can also be used as comonomers. Important for the superabsorbent effect is the generation of charges along the polymer chains when the superabsorbent comes into contact with aqueous liquids. Early superabsorbents were often graft copolymers whose graft base was a polysaccharide, especially starch [43]. Further possibilities are a partly neutralized copolymer of isobutene and maleic acid anhydride [44]; a saponified copolymer of acrylic acid and vinyl acetate [45]; a polyacrylonitrile, grafted onto starch and saponified [46]; a copolymer of acrylamide, 2-acrylamido-2-methylpropanesulfonic acid, and partly neutralized acrylic acid [47]; and a copolymer of acrylamide and trimethylacrylamidopropylammonium chloride [48].

Cross-Linkers. All compounds with at least two polymerizable double bonds, compounds with a polymerizable double bond and a functional group which is reactive towards the monomers, and compounds with at least two functional groups that can react with the monomers are generally suitable as cross-linkers for superabsorbents. Multivalent salts such as aluminum salts can also be used.

Preferably used are highly water soluble compounds such as *N*, *N′*-methylenebisacrylamide, triallylamine, di- or polyacrylic acid esters of polyols, (e.g., low-molecular poly(ethylene glycols) or ethoxylated trimethylolpropane), di- or poly(allyl ethers) of polyols, acrylic acid esters of mono(meth)allyl ethers, and di- or polyglycidyl compounds of polyols or polyamines [49–54].

4.2. Radical Solution Polymerization

4.2.1. Principles [see also (→ Polyacrylamides and Poly(Acrylic Acids, Chap. 3.1.)]

The polymerization rate of acrylic acid in aqueous solution depends on the pH of the solution and consequently on the degree of neutralization of the carboxyl groups [55]. At pH 7 the polymerization rate shows a minimum due to electrostatic repulsion between dissociated charged monomer and charged propagating radicals. Most commercial superabsorbents are sold with a degree of neutralization of between 50 and 85 mol %. Therefore, the low polymerization rate is either accepted or the superabsorbents are patrially neutralized after polymerization. Acrylic acid and the cross-linkers are dissolved in water. Smaller quantities of further co-monomers can be added. The partial neutralization of acrylic acid is normally carried out by addition of sodium hydroxide or sodium carbonate. After removing the oxygen from the solution, the polymerization is typically started with radical initiators. The so-called gel effect [56–59] in free radical polymerization at high monomer concentrations is associated with a slowing down of the rate of termination between active growing chains due to decreasing chain mobility, which leads to autoacceleration and high conversions as long as the momomer can be supplied to the growing chain ends. Due to the exothermic reaction and high reaction rates the gel point of the system is quickly reached, and the temperature rises to 100 °C and higher within minutes. Heat removal is necessary to avoid deterioration of the properties of the final product. The conversion of the polymerization has to be as high as possible, and all side reactions which increase the portion of residual acrylic acid monomer must be avoided.

In the co-polymerization of partially neutralized acrylic acid with cross-linkers bearing multiple vinyl groups, the time at which the cross-linker is incorporated into the growing chains depends on the reactivity ratios. Different kinds of functional groups in cross-linker molecule exhibit different reactivities. When the reactivity of the cross-linker with the monomer is high it is consumed early, so that the polymer chains formed late in the polymerization are less likely to be incorporated into the network. This results in extractable polymers which do not contribute to the modulus of the network. Therefore, cross-linker functionality should be available at the end of the reaction, too. Once the cross-linker molecule is part of the chain, the reaction of the other vinyl groups may be hindered or prevented.

Another possiblity of creating a polyacrylate network is to add substances to the polymerization mixture which can react with the carboxyl groups, such as multifunctional alcohols, glycidyl ethers, or multivalent cations.

Polymer chains that are not connected to the network through cross-linkers can be grafted onto the gel network by grafting reactions, especially at the end of the reaction or during drying. Peroxosulfates can abstract hydrogen atoms and can thus result in further cross-links. Extractable polymers can be formed by routes other than nonincorporation into the network; these include chain degradation and cleavage of cross-links. It is important to generate a homogeneously cross-linked network in which the extractable sol fraction is minimized. Network defects such as intramolecular cyclization, divinyl loops, chains with only one end attached to the network, and unconsumed cross-linker double bonds lower the cross-linking density and affect product quality.

When the polymerization is complete a gel is obtained, which is disintegrated, dried, milled, and sieved to the desired grain sizes.

4.2.2. Production Processes

In practice continuous and semi-continuous production processes are preferred. Acrylic acid, cross-linker, sodium hydroxide, and additives are mixed in storage containers. Consequently the order of the addition can vary. The amount of cross-linkers is between 0.1 and 2 wt % relative to acrylic acid. The degree of neutralization of the acrylic acid is 0 – 80 mol %. If a slightly neutralized acrylic acid monomer solution is polymerized the resulting polymer gel must be treated with an appropriate alkali to achieve the desired degree of neutralization [60]. The aqueous solution is freed from oxygen by purging with nitrogen in a stripper on the way to the reaction vessel or in the reaction vessel itself [61]. The reaction is started with water-soluble initiators. These can be thermal initiators such as peroxosulfates or azo compounds, redox systems such as iron salts and hydrogen peroxide, or photo-initiators. The initiators are metered into the reaction vessel or into the feed stream together with the monomer solution.

During the reaction a large quantity of heat is released. This limits the concentration of the acrylic acid in the monomer solution, which is typically between 15 and 40 wt %. A production process that operates with a high acrylic acid concentration of 50 – 80 % is also known. This process uses the heat of polymerization to evaporate water and gives a dry superabsorbent polymer in a slightly foamed form that has a water content sufficiently low for powdering without any further drying step [62, 63].

Because the throughput must be optimized for an economical process, attention is paid to good heat removal when selecting the polymerization vessels. Furthermore, as little water as possible should have to be vaporized during the subsequent drying step. Therefore, reactions vessels exist in which the gel being generated is disintegrated during the reaction [64, 65]. These kneaders can be equipped with different shafts. Another production process uses a belt [66]. The monomer solution is placed on a belt which is curved upwards in a U-shape at the sides. The initiators are metered into the inflow of monomer solution. After polymerization is completed the generated gel strand is cut into small blocks at the end of the belt [67]. The gel is further disintegrated to ensure optimum drying. Additives can be added to the disintegrated gel. Together with these steps further neutralization of the gel can be performed [68, 69]. In all cases the disintegrated gel is dried to a water content of 5 – 10 wt %. In principal, this can be carried out in all commercial dryers. Circulating belt dryers are preferred [70]. The temperature of the dryers is typically between 100 and 200 °C. The drying step also lowers the residual content of acrylic acid monomer. The acrylic acid evaporates together with the water vapor or reacts with residual initiator.

The dried product is milled and sieved. For use in diapers, normally a grain size between 0.15 and 0.8 mm is desired. Milling is usually carried out in two steps. After coarse milling in an attrition mill, knife mill, or roll crusher, the product is brought to its final grain size in roll crushers. Sieving is performed with tumbler or vibrating screens, whereby the overflow particles return to the milling circuit. The fine fraction is recycled back into the process by adding it to the monomer solution as a powder, as a suspension in a nonswelling solvent [71, 72] or by mixing it with the polymerized gel particles and water [73].

4.3. Inverse Suspension Polymerization

Superabsorbents produced by inverse suspension polymerization [see also → Polyacrylamide and Poly(Acrylic Acids), Chap. 3.2.2.] represent only a minor portion of the superabsorbents used commercially in diapers. Nevertheless, they have advantages and are predominantly used in technical applications such as cables.

The aqueous monomeric solution is prepared in the same way as for radical solution polymerization, but polymerization is then performed by dispersing the monomeric solution in a solvent which is inmiscible with the monomer solution, e.g., $C_6 - C_{14}$ aliphatic hydrocarbons. The initiator is soluble in the monomer solution and is usually added to it before dispersion. The polymerization is performed by heating the dispersion. To avoid coalescence and sticking together of the polymer particles during and after polymerization, a stabilizer is added to the continuous hydrocarbon phase before the suspension

is formed. Depending on the desired size of the particles, stabilizers can be nonionic surfactants such as sorbitan monostearate [74], polymers such as ethyl cellulose or oxidized polyethylene [75], as well as particulate suspension aids such as cationic bentonites [76]. The suspension aid used and the shape and rotating speed of the stirrer affect the shape and size of the particles. As a rule round particles are generated. By appropriate reaction control, however, agglomerated, porous, elongated, and convoluted particles can also be made [77,78]. Reprocessing is carried out either by filtration with subsequent drying or by azeotropic distillation to remove the water. Residual water and the continuous phase must finally be separated by drying.

Advantages of this method are a better heat dissipation during polymerization, the omission of gel disintegration and milling, and the possibility of modifying the surface of the particles directly afterwards in the same reactor. Disadvantages are that the continuous-phase solvents must be recycled. This is expensive, as they are volatile and highly flammable substances. Furthermore, the reactor dimensions are limited by the required stability of the suspension.

4.4. Surface Treatment

In the development of thinner diapers the superabsorbent partly had to take over the functions fulfilled by the cellulose fluff. The liquid distribution inside the diaper is hereby of central significance for using of the entire absorption capacity of the diaper. One problem with early superabsorbents was the tendency to clump when aqueous liquids were added. The liquid was hindered from penetrating further into the diaper, the diaper leaked, and the baby became wet. The gel particles were were soft and easily deformable, and the ability to swell under an applied load was low. For this reason modern superabsorbents for diapers are modified so that cross-linkage density at the surface is higher than inside the particles [79]. The corresponding test and the effects on swelling capacity and modulus of the gel are discussed in Chapter 5.

There are different ways to achieve a gradient of cross-linking density. Subsequent cross-linking of the surface with substances which can react with the carboxyl groups of the polyacrylic acid is mainly used. For example di- or polyalcohols [80], di- or polyglycidyl compounds [81], alkylene carbonates [82], or trivalent ions like aluminum ions [83] are used. The reaction is performed after spraying the substances onto the superabsorbent particles by a heat treatment at 90 – 250 °C in fluidized-bed, infrared, paddle [84], rotating-disk [85], or vortex dryers. A core – shell particle is obtained which, depending on the thickness of the shell, is vulnerable to attrition during handling.

A different method to achieve more extensive cross-linking of the surface is the addition of an oxidizing agent such as sodium chlorate to the polymerization, which during subsequent heating generates a gradient of cross-linking density from the inside to the outside [86]. In this case high temperatures are necessary, which can be detrimental to other properties.

5. Properties, Characterization, and Analysis

The demanded or required characteristics of superabsorbents depend very strongly on the application. For example, long-term stability in the swelled state or the ability to swell repeatedly is of minor significance for use in diapers, but for application in soils it is decisive. This chapter presents the fundamental tests for characterizing superabsorbents for hygiene products [87]. Particular requirements for superabsorbents for other applications are discussed in Chapter 6.

A superabsorbent is mainly characterized by its swellability, which depends on the structure of the network, that is, the cross-linking density, the modulus of the network, the charge quantity and density, and the completeness of integration of the monomers and the polymer chains in the network. As most superabsorbents are granules, characteristics of the bulk granules such as bulk density, particle-size distribution, shape, and surface constitution are of great significance in all tests. All tests are carried out with a large number of particles, as measurements on the individual grains are very expensive and seldom representative. This also means that the test results depend on the particle characteristics.

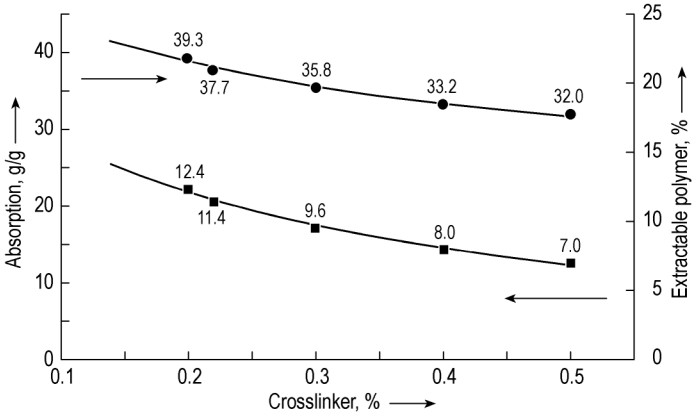

Figure 2. Centrifuge Retention Capacity (ERT 441.1-99) and extractable polymer (ERT 470.1-99) versus amount of cross-linker. The amount cross-linker relates to the content of acrylic acid.

5.1. Absorption Capacity

The absorption capacity of superabsorbents granules mainly depends on the density and homogeneity of the cross-linking of the polymer network, the number of charges on the polymer chains (degree of neutralization), and the ionic strength of the swelling. The higher the cross-linking density the lower the absorption and the lower the amount of extractable polymer (Fig. 2). The absorption capacity is usually determined gravimetricaly.

Ionic effects have a strong influence on the absorption capacity. The higher the charge on the polymer chain (i.e., degree of neutralization) and the lower the ionic strength of the swelling liquid, the higher is the absorption (Figs. 3 and 4). Trivalent cations reduce the absorption more than bivalent cations, which in turn are more effective than monovalent cations. As substitute for urine either a synthetic urine solution or, more simply a 0.9 % salt solution is used. The way in which the test is performed also affects the determined degree of absorption. As the test system consists of superabsorbent granules, a part of the liquid is not fixed inside the superabsorbent particle itself but between the individual gel particles (interparticle or interstitial liquid). If this liquid is removed by centrifuging or brief application of a vacuum, the fluid retention of the superabsorbent itself is obtained.

A standard test for the absorption capacity is the centrifuged absorption capacity CRC (ERT 441.1-99) which is the ratio of the absorbed liq-

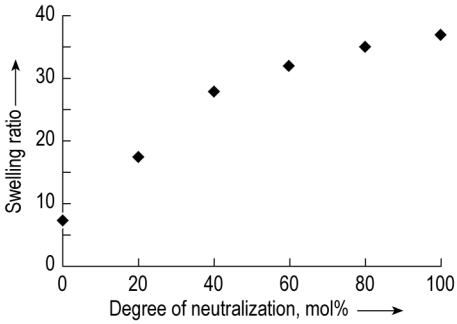

Figure 3. Swelling ratio versus degree of neutralization of a partially neutralized polyacrylate gel. The gel was neutralized after polymerization and swollen in a NaCl solution with a concentration of 153.8 mmol/L. Taken from the Y. Yin, Swelling Behavior of Polyelectrolyte Gels, Ph. D. Thesis, Princeton University 1993, p. 130.

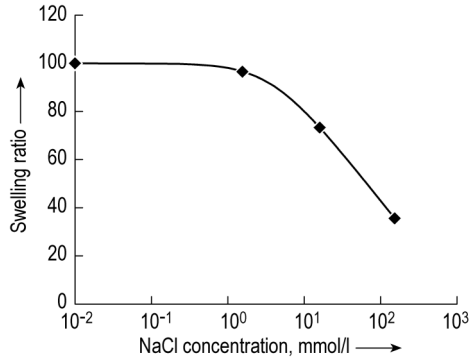

Figure 4. Swelling ratio versus NaCl concentration of the solution for a partially neutralized polyacrylate gel. The degree of neutralization was 80 %. Taken from the Y. Yin, Swelling Behavior of Polyelectrolyte Gels, Ph. D. Thesis, Princeton University 1993, p. 130.

uid to the mass of the dry superabsorbent. 0.2 g of superabsorbent granules are filled into a heat-sealable teabag, which is subsequently sealed. The teabag is weighed and then immersed in of 0.9 % NaCl solution for 30 min. Then it is transferred to a centrifuge (Fig. 5) and spun at 1600 rpm for 3 min, after which it is weighed again. The same procedure is carried out on a teabag without superabsorbent as a blank value. The retention is calculated from the weigh of the filled teabag after swelling and centrifugation m, the weight of the dry superabsorbent filled into the teabag m_s, and the and weigh of the empty teabag after swelling and centrifugation m_t.

$$\mathrm{CRC} = \frac{m - m_s - m_t}{m_s} \tag{12}$$

5.2. Shear Modulus

The viscoelastic behavior of a substance is measured by subjecting it to a harmonically oscillating shear deformation γ. The shear modulus G is determined from the shear stress σ. The shear modulus G can be defined as a complex quantity [88]:

$$G = G\prime + iG'' \tag{13}$$

where G' is the storage modulus and is a measure for the elastically stored energy, whilst the loss modulus G'' represents the energy which is dissipated by viscous flow.

A definition of a gel in terms of dynamic mechanical properties is that the storage modulus G' exhibits a pronounced plateau and the loss modulus G'' is considerably smaller than the storage modulus in the plateau region [89]. G' and G'' are measured with an oscillatory rheometer. Normally a plate – plate geometry is used. In practice, the pre-swelled superabsorbent granule is used. A defined grain fraction of the superabsorbent granule is pre-swelled to 70 – 90 % of the maximum swelling in 0.9 % NaCl solution. The pre-swelled granule is positioned on the lower plate. The gap between the plates is adjusted to the size of the gel particles. After lowering the upper plate, the gel particles which project over the edge of the plates are removed. An oscillating strain with a frequency of, e.g., 1 Hz is applied to one of the plates.

The swollen particles between the plates are in contact between each other. The shear modulus depends on the binding forces and the mobility between particles, and not only on the network structure of the gel. Therefore, the strain amplitude has to be very small, or measurements are carried out at different amplitudes and extrapolated to zero strain. Even then the method only gives meaningful results when the cross-linking density of the particles is not too low or high and when the particles are not surface-cross-linked. A more appropriate method is the use of a pre-formed disk of gel which fits perfectly into the gap between the plates of the rheometer. Such measurements [90] showed that in accordance with theory, the storage modulus increases with increasing cross-linker concentration (Fig. 6) or efficiency and with increasing monomer concentration in production. The

Figure 5. Centrifuge Retention Capacity (ERT 441.1-99)

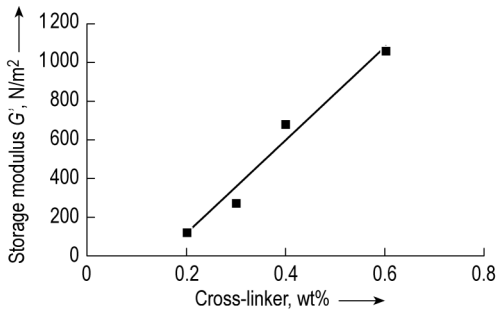

Figure 6. Storage modulus (frequency 1 Hz, amplitude 1 mrad) of sodium polyacrylate gels as a function of the amount of cross-linker used in polymerization. 1 g superabsorbent powder (300 – 600 µm) was swollen in 50 mL of deionized water.

as-prepared degree of neutralization also influences the storage modulus but the direction is controlled by the cross-linker density.

5.3. Kinetics of Absorption

The rate at which a liquid is absorbed into the particle can be determined by measuring absorption versus time. The time required for a specified percentage of the maximum absorption to be reached is a measure of the rate.

In this test the superabsorbent must absorb the liquid from a reservoir. The data can be fit to a simple exponential [91]

$$\frac{\mathrm{d}q}{\mathrm{d}t} = k\left[q_{\max} - q\left(t\right)\right] \tag{14}$$

where $q_{\max}$und $q(t)$ are the absorption capacities at equilibrium and at any time t under the conditions of the experiment. The rate constant k depends on the diffusion coefficient of the liquid into the particle D and the radius of the particle r.

$$k = \frac{\pi^2 D}{r^2} \tag{15}$$

The diffusion coefficient depends on the temperature and on the salt concentration of the liquid.

Other test methods determine the time which a specified amount of superabsorbent requires to absorb a certain quantity of liquid after mixing (vortex test [92], dynamic swelling rate [93]).

The absorption rate can be improved by increasing the porosity of the particles and thereby increasing the contact area. This can also be achieved by changing the shape. Superabsorbent fibers exhibit very fast absorption. Since the absorption rate is higher for smaller particle, the formation of agglomerates of smaller particles increases the rate of absorption without changing the particle-size distribution.

The wettability of the particle surface influences the initial absorption. A hydrophobic coating delays the uptake of liquid. Since the tests are carried out with a large number of particles, properties such as the permeability of the granuler mass (or bulk) are important. This is influenced by the packing of the particles, which in turn is influenced by the shape and the particle-size distribution of the particles.

5.4. Swelling under Load

The requirement for superabsorbents to absorb liquids under external pressure resulted from their application in diapers, for which improved performance was observed for superabsorbent which were less easily deformed and deswelled by the weight of a baby. The magnitude of the deformation depends on the modulus, which in turn mainly depends on the effective cross-link density and the extent of swelling.

Using scaling theory Buchholz derived an equation [5] in which the swelling capacities Q_{e} and Q of a gel in the unloaded and loaded state are proportional to the modulus of the fully swollen gel G_{e} and the applied pressure P.

$$\left(\frac{Q_{\mathrm{e}}}{Q}\right)^{9/4} - \left(\frac{Q_{\mathrm{e}}}{Q}\right)^{1/3} \approx \frac{P}{G_{\mathrm{e}}} \tag{16}$$

Q is defined as the ratio of the gel mass to the dry mass of the polymer. The absorption capacity CRC used in industrial practice is the ratio of the liquid mass absorbed to the dry polymer mass. The equation states that at a given modulus G_{e} and therefore a given swelling capacity Q_{e} of the fully swollen gel, the swelling capacity in the loaded state is lower at higher loads. On the other hand, the magnitude of the deformation depends on the modulus, which in turn mainly depends on the effective cross-link density and the extent of swelling. Highly cross-linked particles with low absorption capacity are able to swell sufficiently under an external load. By modifying the surface to give a core – shell particle with a higher cross-linking density near the surface, superabsorbents particles could be obtained that combine absorption capacities both with and without an applied load (Fig. 7, see next page).

For the determination of the absorption under load (AUL [94]) or absorption against pressure (AAP [95]) a measuring device is used that consists of an acrylic glass cylinder which is closed at the bottom with a metal gauze (Fig. 8, see next page). A specified amount of superabsorbent granules is filled into this cylinder. The measuring cylinder can be loaded with different weights. Usually loads of 0.3, 0.7, and 0.9 psi (ca. 2.1, 4.9, and 6.3 kg/m^2) are used. The absorption of the liquid is performed by placing the cylinder onto a porous ceramic filter plate, which is covered with filter paper. The ceramic

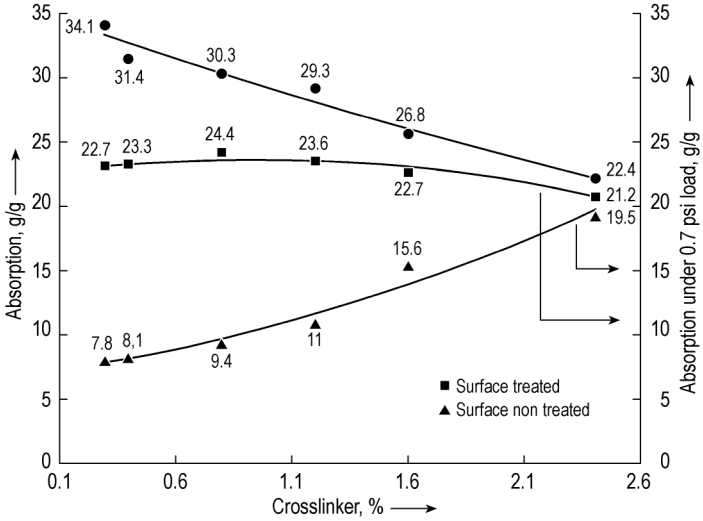

Figure 7. Centrifuge Retention Capacity (ERT 441.1-99) and absorption against pressure (ERT 442.1-99, 0.7 psi load) as a function of the amount of cross-linker. The amount of cross-linker is related to the content of acrylic acid. The circles represent the centrifuge retention capacity of the samples, the triangles represent the absorption against the pressure of the samples and the squares represent the absorption against pressure of the surface treated sample.
0.7 psi $\approx$ 4.9 kg/m^2.

Figure 8. Absorption Against Pressure (ERT 442.1-99, 0.7 psi load).
0.7 psi $\approx$ 4.9 kg/m^2.

filter plate lies in a tray which is filled to its upper edge with 0.9 % NaCl solution. After 1 h the cylinder is weighed again. The absorption under pressure is calculated from the difference to the initial weight and the weigh-in of the superabsorbent, including the liquid in the pores between the particles. The resulting absorption capacity under load is also dependent on the way in which the liquid is transported in the particle bed. The liquid can be transported through the galled particle bed by capillary forces as long as the porosity of the particle bed is high enough and the pores between the particles are connected. When the particles absorb liquid and start swelling, they soften and can be deformed more easily. Hence, the modulus of the gel network decreases with increasing liquid absorption. An applied load deforms the particles and blocks the pores to a certain extent. This restricts further liquid transport to diffusion through the particles themselves which is a much slower process. A high cross-linking density, especially near the surface, helps to prevent deformation.

5.5. Permeability

The distribution of the liquid in the diaper, that is, the transport of the liquid from the point of entry to the rest of the diaper is important to fully use the absorption capacity of the diaper. Modern diapers contain large amounts of superabsorbents. Therefore the swollen superabsorbent particle mass has to be permeable for additional liquid to be transported through the gel mass, particularly when a load is applied to the gel. Low permeablity means blockage of the gel for further liquid, and the diaper may leak. Permeability depends on the particle-size distribution, the free volume of the pores and the connection

between them, and the ability of the individual gel particles to maintain their shape in the swelled state under pressure. Permeability is a function of the porosity, geometry, packing, and modulus of the gel particles. The permeability is measured by swelling the superabsorbent granules under pressure in an acrylic glass cylinder with a bottom of metal gauze. The throughput quantity per unit time of the liquid through the gel under load is a measure of the permeability [96].

5.6. Test Methods

Test methods for superabsorbents recommended by the European Disposables and Nonwovens Association (EDANA) are listed in Table 1.

5.7. Stability of Superabsorbents

The thermal stability of superabsorbent polymers depends on the degree of neutralization. Poly(acrylic acids) undergo dehydration commencing at about 150 °C on programmed heating at 10 K/min, predominantly by formation of an intramolecular anhydride between neighboring pendent carboxyl groups. Decarboxylation of the anhydride becomes significant from about 250 °C, and major fragmentation of the modified backbone occurs above 300 °C [97]. At degrees of neutralization exceeding 50 mol % evaporation of tightly bound water, rather than anhydride formation, is the assumed route for weight loss at lower temperatures [98]. Fully neutralized sodium polyacrylates are thermally stable polymers that undergo degradation only above 380 °C. The volatile degradation products are mainly carbon dioxide and acetone. Small amounts of alkenes, ketene, higher ketones, and aromatic hydrocarbons are also formed. The residue consists mainly of sodium carbonate together with some carbon. Polyacrylates neutralized with other metals, especially transition metals, can show different degradation behavior [94]. Once swollen, the stability of the gel is significantly lower. Mechanical stress, heat, UV radiation, traces of radicals or metal ions like Fe^{2+}, Fe^{3+}, Cu^{2+} or Al^{3+} in combination with oxygen result in a breakdown of the gel structure and the increase of the water-soluble fraction within a short time.

6. Uses

6.1. Personal Hygiene Products

The main application area for superabsorbents is personal hygiene: infant diapers, adult care articles, and feminine sanitary napkins. More than 95 % of the superabsorbents are consumed in this area.

Table 1. EDANA recommended test methods (ERT methods)

Test method	Description	Reference
pH	pHh value of a superabsorbent dispersion	ERT 400.1-99
Residual monomers	content of extractable acrylic acid, HPLC	ERT 410.1-99
Particle size distribution	screens with different mesh sizes	ERT 420.1-99
Moisture content	weight reduction after 3 h at 105 °C	ERT 430.1-99
Free swell capacity	free swell capacity in saline solution	ERT 440.1-99
Centrifuge retention capacity	fluid retention capacity in saline solution following centrifugation	ERT 441.1-99
Absorption against pressure	determination of the capacity to absorb saline solution under a specified pressure	ERT 442.1-99
Flow rate	time for a given mass of superabsorbent powder to pass through a specified funnel	ERT 450.1-99
Flowability	mass of powder (g) per unit time (s) flowing through a specified funnel	
Density	determined by pouring a representative sample through a specified funnel into a density cup	ERT 460.1-99
Extractables	sum of the acid and salt groups of monomeric, oligomeric, and polymeric carboxylates extracted from the superabsorbent polymer in saline solution	ERT 470.1-99
Respirable particles	particles having a diameter of less than 10 µm, determined by using a laser light-scattering particle size analyzer	ERT 480.1-99

6.1.1. Diapers

Modern diapers can barely be compared with the fabric diapers of 30 years ago. Especially in industrial nations, the disposable diaper has become standard.

A typical modern diaper (Fig. 9) consists of a permeable cover layer (coverstock), a distribution system (distribution layer), an absorption layer (absorbent core) and an impermeable outer layer (impermeable back sheet).

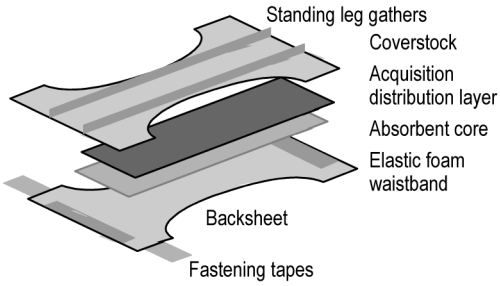

Figure 9. Schematic of a baby diaper

The cover and distribution layers have the task of taking up the baby's urine and distributing it over a wide area of the absorbent core. These webs are usually hydrophilically treated nonwoven fabrics mainly made of propylene. The absorbent core should store the liquid without releasing it to the surface of the diaper, in order to ensure a dry diaper. In the simplest case the outer layer of the diaper is a water-impermeable polyethylene foil. Improved foils are air-permeable.

The absorption core consists of a mixture of cellulose fluff and superabsorbent granules. The cellulose fibers provide for good distribution of the liquid in the entire layer but can not effectively store liquid themselves. While in the past the fluff percentage was at 80 – 90 %, today it is only ca. 50 % (Fig. 10). Thus diapers became thinner and the density of the core was increased. When early superabsorbents were developed the main emphasis was on absorption capacity. Due to the low modulus of the network, they were soft and could be easily deformed. They tended to block the liquid at the point of entry, and this led to leakage of the diaper. The small amounts of fluff used make this negative effect even more pronounced. The increased amount of superabsorbents in the absorbent core required improved superabsorbent properties. Even with an overlying distribution layer the ability of the absorbent core to distribute the liquid throughout the entire core is essential for using the total absorption capacity. The cross-linking density and the gel strength were increased, absorption capacity was decreased. Superabsorbents with special properties, such as the ability to swell under an applied load, low or high absorption rates, or unusual particle sizes, led to improved diaper designs. Fast-running diaper machines and modern storage and conveying systems demand an adjusted flowability. Health and safety regulations require a low content of dust.

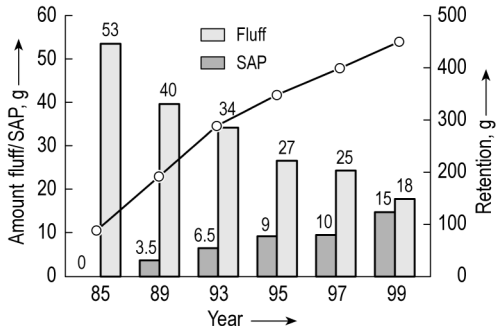

Figure 10. Change of fluff/SAP ratio in diapers since 1985

Diapers for the adult care area distinguish themselves in particular by the absorption capacity, which must be greater. The portion of cellulose fibers is also higher. Special superabsorbents were also developed for adult care articles. For example, superabsorbents with improved ability to reduce odors [99] are used in this area.

6.1.2. Feminine Hygiene Products

Feminine sanitary napkins mainly consist of a cover, an absorption layer, in which superabsorbents can ce incorporated, and an outer layer (coverstock, absorbent core, back sheet). Frequently pre-manufactured products, such as for example airlaid composites, into which the superabsorbent is incorporated, are used as absorption layers. Feminine napkins have to be very thin. Therefore, the particle size of the superabsorbent granules used is normally smaller than in diapers. Superabsorbent fibers are applied as

well. Menstrual liquids are more difficult to absorb than urine. They are viscous fluids consisting of water, salts, proteins, and blood cells. Protein and blood cells are too large to be absorbed by the superabsorbents. They are only adsorbed on the surface of the superabsorbent particles where they impede the absorption of water and dissolved salts.

6.2. Packaging Applications

The ability of superabsorbents to retain aqueous liquids has also found applications in food packaging systems. Prepacked meat and poultry packages have an absorbent pad containing superabsorbent granules or fibers which is integrated into the tray or glued to the floor of the tray. So-called soaker pads are also used during the transport of fish. The absorbent pad must be able to absorb all of the water when the ice used for cooling melts. The absorbing pads typically consist of an inner part and a coverstock. The inner part can be a mixture of cellulose fibers and superabsorbents, an air-laid composite, or a laminate containing nonwoven fabrics and superabsorbent granules or fibers. On the underside the cover can consist of a perforated foil or a hydrophilic fabric, and the upper side of a liquid-impermeable foil or a hydrophilic fabric. All components of these absorbing pads must be approved for the food use [100]. This also results in particular limitations for the chemical composition of the superabsorbents.

In addition superabsorbents are used in smaller quantities as refrigerants. The swollen gel is frozen as a laminate or directly packed in plastic bags. Superabsorbents are also used to absorb condensation water in packaging or thicken contaminated liquids in transport containers in hospitals.

6.3. Soil Conditioners

Horticulture, agriculture, and forestry are a wide application field for superabsorbents. Due to their ability to store water and nutrients for uptake by plants they are used as soil conditioners. Examples of applications in the horticulture area are the use in flowerpots for ornamental plants, the integration into golf courses or football grounds, landscaping of embankments, roof gardens, and noise barriers, as well as the use in planting holes for trees and bushes. In agriculture the superabsorbents are applied in seed and plant furrows to increase the water-use efficiency. The best results are achieved on sandy soils in arid areas, where the ability to store water and nutrients has the most positive effect on water management and growth. Yield increases could be achieved for almost all useful plants, such as wheat, corn, alfalfa, potatoes, asparagus, and tomatoes.

In reforestation superabsorbents provide for better root-taking of the young trees, a lower mortality rate, and therefore lower planting costs.

Applications in soils create special requirements for superabsorbents. They must swell and shrink under the soil conditions and remain stable for several years. Since soils contain large amounts of electrolytes, superabsorbents are required which are effective at high salt concentrations and in the presence of multivalent cations. Copolymers made of acrylamide and acrylic acid are mainly used.

A particularly interesting application is the possibility to store biologically active ingredients such as plant-protecting agents in superabsorbents and to apply them in this form [101, 102]. The controlled release of the active ingredients reduces the required amount of the agent by extending the duration of action. The number of applications can be reduced, harmful side effects can be diminished, and new possibilities of application for existing plant protectives can be created.

6.4. Cable Applications

Superabsorbents are used in cables to avoid the penetration of water into the cable when the cable sheath is damaged. Mostly superabsorbents are used which swell very quickly and prevent the water spreading inside the cable by forming a water-impermeable gel layer. Furthermore, this gel layer must be stable over a certain period of time. For applications in cables laid on the seabed, superabsorbents are required

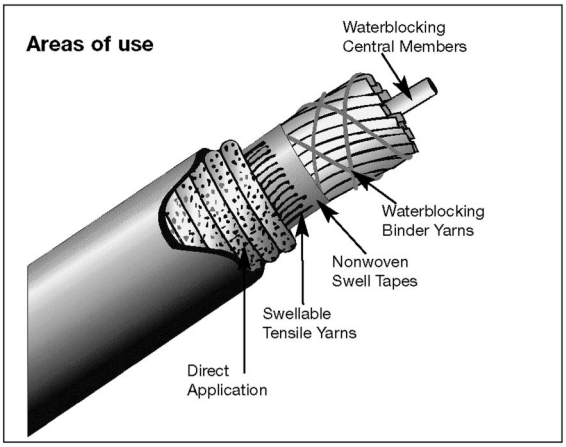

Figure 11. Depiction of where superabsorbents can be used in a cable

which are insensitive to bivalent cations. Polymers with other monomers than acrylic acid are used, as well, i.e., copolymers of maleic anhydride and isobuthylene. Particle sizes between 45 and 300 µm are mostly used. The particles are normally fabricated in to water-blocking tapes [103, 104] by mixing the powder with a polymeric binder and applying it to a nonwoven fabric. The particles can also be incorporated directly into the gaps of the cable during production (Fig. 11) [105].

6.5. Civil Engineering and Construction

In the building industry superabsorbents are used in sealing materials. For example, superabsorbents made from maleic anhydride and isobutylene are incorporated in rubber sealing tapes, which are used in joints between concrete blocks to seal against water. Superabsorbents can be added to concrete, cement, mortar, or ceramics to achieve improved product characteristics such as extended setting time, lower water stress, or increased porosity. Until now, superabsorbents have only been used in specially products. There are numerous possible applications during the conditioning and thickening of mud. Positive results were obtained with secondary scale [106], paint sludge [107], and garbage-incineration ashes [108]. Industrial-scale implementation is currently hindered for economic reasons.

6.6. Other Applications

In medicine, small quantities of superabsorbents are used in wound dressings to improve healing in humid environments and in sticking plasters [81]. Superabsorbents are used for pet care as pet sheets and kitty litter. Other applications are in perspiration absorption in belts or gloves, as absorbents in portable toilets, as gelling agent in deodorants, and as additives in lotions. Hydrogels also act as sensors. They can swell or shrink spontaneously in response to changes in temperature, pH, or electric fields. They are even used as artificial snow. There is a hall in Tokyo where skiing is possible on snow made of frozen superabsorbent gels. Further applications are continually being discovered, so that this list cannot be complete.

7. Other Forms of Superabsorbent Polymers

Although superabsorbents are currently offered almost exclusively in the form of granules or powders, other forms such as fibers and foams can also be produced. Due to their characteristics and production costs their application is as yet limited.

7.1. Fibers

Superabsorbent fibers are an interesting alternative to granules. Due to their shape they can easily be processed with other fibers. The absorption rate is improved through the higher ratio of surface to volume. Disadvantages are the low mechanical stability, the sensitivity to humidity, their low bulk density, a low swelling ability under pressure, and the high production costs. Besides hygiene products, they are mainly used in technical products such as cables, filters, and packaging.

They are typically manufactured by spinning polymer solutions containing a cross-linker, whereby cross-linking is performed during drying. The fibers are subsequently cut. The polymers can consist, for example, of acrylic acid, an alkyl acrylate as softener, and a self-cross-linking monomer such as hydroxyethyl acrylate [110]. Cross-linking is performed during drying. Copolymers made of isobutene and maleic anhydride [111] with polyols as cross-linkers are also used. Another possibility is the surface modification of nonabsorbent fibers such as polyacrylonitrile fibers [112]. The addition of sodium hydroxide generates carboxylate groups, which impart absorbent characteristics on the fiber.

7.2. Foams and Films

Films of pure superabsorbent currently have no commercial significance. Polymer solutions or emulsions are available, which are applied to fibers or textile surface structures [113, 114]. The polymers are either self-cross-linking by heating or exposure to UV light, or a cross-linker is added. These polymers are based on the same monomers which are used for fibers or granules.

There are two different types of superabsorbent foams. First, the foam itself can consist of superabsorbent material, produced by foaming a monomer solution with subsequent polymerization [115, 116] or by thermal foaming of a polymer solution or emulsion, which can contain cross-linker, blowing agents, and other additives [117, 118].

Second, superabsorbent granules can be integrated into a foamed matrix. The foamed matrix is used for fixing the superabsorbent granule and distributing the liquid, but does not contribute greatly to the liquid absorption. The matrix can be, for example, a polyurethane foam [119]. Latex foams are also used [120, 121].

Special foams are known under the name HIPE (high internal phase-ratio emulsion) [122] which have very good absorbency characteristics, but strictly speaking are not superabsorbent. Their absorbency effect is based on the creation of a very dense capillary network. The monomers form the continuous oil phase of a water-in-oil emulsion. The aqueous phase is used as a place holder for the pores and is removed after polymerization by washing and drying. An open pore structure of low density is thus created. As hydrophobic monomers are mostly used, the surface must be chemically modified by introducing polar groups [123, 124] to enhance the wettability of the foam structure for aqueous liquids.

8. Environmental Aspects and Toxicology

Cross-linked polymers of partially neutralized acrylic acid are the main superabsorbent polymers. Therefore, they were tested in various systems to elucidate their toxicological profile both for handling during production and use in articles such as diapers [125, 126].

They are essentially insoluble in water, but they contain up to 10 wt % of water-extractable fraction and up to 500 ppm of residual sodium acrylate.

Superabsorbent polymers based on partially neutralized acrylic acid exhibit a toxicological profile of low concern. Under appropriate test conditions there were no signs of acute oral or dermal toxicity. Furthermore subacute oral toxicity and long-term dermal toxicity have not been observed in rats [127]. The eye-irritating potency (rabbit) seems very low. Absorption after oral uptake is negligible [128]. They show no evidence of allergic contact sensitization in guinea pigs and humans and no irritating properties. No mutagenic and teratogenic potency was found [129]. In a chronic inhalation study in rats, micronized fine dust of a superabsorbent polymer caused tumor formation in rat lungs at the highest concentration. This effect is attributed to a chronic inflammatory process and not to a direct genotoxic action [130]. However, the use in

consumer products and technical articles must be regarded as safe.

Superabsorbent polymers will eventually enter the environment via different routes and in various forms like exhaust gas after incineration of household waste (diapers, etc.), compost, household waste in landfills, or biosludge, which is incinerated, landfilled, or used in agriculture. Taking realistic predicted environmental concentrations and the ecotoxicity data into account there is no concern for the environment.

Pollution of ground water is highly unlikely as extracts strongly bind to the soil matrix with only little desorption and subsequent leaching of a very small mobile fraction of low molecular mass which seems to be potentially biodegradable [131].

Biodegradability studies in aquatic systems show that superabsorbent polymers based on partially neutralized acrylic acid are poorly biodegradable. In soil there is a potential for biodegradation once the polymer is water-soluble. In an optimized simulation of an intensive aerobic composting process a superabsorbent polymer was degraded by 14 % after 45 d. The metabolism of the aqueous soluble fraction of a superabsorbent polymer in soil was monitored under aerobic conditions. 20 % degradation to CO_2 resulted after 39 weeks. In a recent study white-rot fungus could both solubilize and mineralize the polymer in the soil. Soil microbes alone were not able to degrade the polymer but cooperated with the fungus during polymer degradation in soil [132, 133].

9. Economic Aspects

Superabsorbents are used almost exclusively in personal hygiene products. World consumption in 2001 was 1 100 000 t. Of this 83.5 % went into diapers, approximately 8.5 % into adult care articles, 5 % into feminine sanitary napkins, and 3 % into other applications. The industry is concentrated in a few companies, both in personal hygiene products and superabsorbents. There are two major manufacturers of diapers (Kimberly-Clark, Procter & Gamble) and four major manufacturers of superabsorbents (BASF, Nippon Shokubai, Dow, Degussa).

Worldwide market growth is estimated at 5 % per year, whereas for North America, Europe, and Japan no market growth is assessed. In 1995 [134] consumption was only ca. 700 000 t/a. The rapid increase since then can be explained mainly by the increased amount of superabsorbent in diapers, which made the development of thinner diapers possible. Diapers reduced in volume not only have esthetic appeal, they also make economical sense. Besides production costs, packing, and transport costs are reduced. Whether the quantities will increase any further is not yet clear. The future increase in superabsorbent production for the personal hygiene market is certainly connected to the growth of the markets for diapers in Eastern Europe, South America, and Asia. However, at the same time, the use of superabsorbents for technical applications is only just beginning.

10. References

1. Y. Hu, Z. Zhang, B. He, *Lizi Jiaohuan Yu Xifu* **15** (1999) 378 – 384.
2. S. Kawamura, M. Okikura, *Porima Daijesuto* **43** (1991) 71 – 82.
3. M. H. Zhou, C.-S. Ha, W.-J. Cho, *J. Appl. Polym. Sci.* **81** (2001) 1277 – 1285.
4. J. Croquette, C. Bocard, *Oil Petrochem. Pollut.* **1** (1983) 261 – 267.
5. F. L. Buchholz, A. T. Graham: *Modern Superabsorbent Polymer Technology,* Wiley, Chichester 1998.
6. L. Brannon-Peppas, R. S. Harland: *Absorbent Polymer Technology – Studies in Polymer Science,* Elsevier, Amsterdam 1990.
7. J. R. Gross in L. Brannon-Peppas, R. S. Harland (eds.): *Absorbent Polymer Technology – Studies in Polymer Science,* Elsevier, Amsterdam 1990.
8. P. K. Chatterjee: *Absorbency,* Elsevier, Amsterdam 1985.
9. Y. L. Yin, Sweling Behavior of Polyelectrolyte Gels, Ph. D. Dissertation, Princeton University, 1993.
10. R. S. Harland, R. K. Prud'homme (eds.): *Polyelectrolyte Gels: Properties, Preparation and Application,* American Chemical Society, Washington, D.C.1992.
11. F. L. Buchholz, A. T. Graham: *Modern Superabsorbent Polymer Technology,* Wiley, Chichester 1998.
12. U. Brendel, Dissertation, Mainz 1999.
13. P. J. Flory: *Principles of Polymer Chemistry,* Cornwell University Press, Ithaka 1953.

14. H. M. James, E. Guth, *J. Chem. Phys.* **15** (1947) 669.
15. J. E. Mark, B. Erman: *Rubberlike Elasticity,* Oxford University Press 1975.
16. P. J. Flory, B. Erman, *Macromolecules* **15** (1982) 800 – 806.
17. P. J. Flory, B. Erman, *Macromolecules* **15** (1982) 806 – 811.
18. U. P. Schröder, W. Oppermann in J. P. Cohen Addad (ed.): *Physical Properties of Polymeric Gels,* Wiley, Chichester 1996.
19. U. Brendel, Dissertation, Mainz 1999.
20. P. J. Flory: *Principles of Polymer Chemistry,* Cornwell University Press, Ithaka 1953.
21. H. H. Hooper, H. W. Blanch, J. M. Prausnitz in L. Brannon-Peppas, R. S. Harland (eds.): *Absorbent Polymer Technology – Studies in Polymer Science,* Elsevier, Amsterdam 1990.
22. U. P. Schröder, W. Oppermann in J. P. Cohen Addad (ed.): *Physical Properties of Polymeric Gels,* Wiley, Chichester 1996.
23. Y. L. Yin, Sweling Behavior of Polyelectrolyte Gels, Ph. D. Dissertation, Princeton University, 1993.
24. M. Zrinyi, F. Horkay, *Polym. Bull.* **3** (1980) 665 – 670.
25. F. Horkay, M. Zrinyi, *Polym. Bull.* **4** (1981) 361 – 368.
26. F. Horkay, M. Zrinyi, *Macromolecules* **15** (1982) 1306 – 1310.
27. M. Rubinstein, R. C. Colby, A. V. Dobrynin, J.-F. Joanny, *Macromolecules* **29** (1996) 398 – 406.
28. F. L. Buchholz, A. T. Graham: *Modern Superabsorbent Polymer Technology,* Chapter 4.6.1., Wiley, Chichester 1998.
29. T. Tanaka, *Phys. Rev. Lett.* **40** (1978) 820.
30. M. Ivalsky, *Macromolecules* **15** (1982) 782.
31. T. Tanaka, I. Nishio, S.-T. Sun, S. Ueno-Nishio, *Science* **218** (1982) 467.
32. I. Ohmime, T. Tanaka, *J. Chem. Phys.* **11** (1982) 5725.
33. Y. Nishiyama, M. Satoh, *Macromol. Rapid Commun.* **21** (2000) 174 – 177.
34. T. Tanaka, D. J. Fillmore, S.-T. Sun, I. Nishio, G. Swislow, A. Shah, *Phys. Rev. Lett.* **45** (1980) 1636.
35. A. Khokhlov, E. Y. Kramarenko, *Macromol. Theory Simul.* **3** (1994) 45.
36. T. Tanaka, D. J. Fillmore, S.-T. Sun, I. Nishio, G. Swislow, A. Shah, *Phys. Rev. Lett.* **45** (1980) 1636.
37. N. Ohmime, T. Tanaka, *J. Chem. Phys.* **11** (1982) 5725.
38. M. Shibayama, T. Tanaka, *Adv. Polym. Sci.* **109** (1993) 1 – 62.
39. Procter & Gamble, 32,649 1988 (K. Brandt, S. Goldman, T. Inglin).
40. R. Pø, *J. Macromol. Sci. Rev. Macromol. Chem. Phys.* **C 34** (1994) no. 4, 607 – 682.
41. A. M. Mathur, S. K. Moorjani, A. B. Scranton, *J. Macromol. Sci. Rev. Macromol. Chem. Phys.* **C 36** (1996) no. 2, 405 – 430.
42. Mark et. al.: *Encyclopedia of Polymer Science and Engineering,* **Vol. 7,** 1988, pp. 783 – 807.
43. Sanyo Chemical Ind. Ltd., US 4 076 663, 1976 (F. Masuda, K. Nishida, A. Nakamura).
44. Dow Chemical Co., US 3 983 095, 1973 (R. N. Bashaw, B. L. Atkins).
45. Sumitomo Chemical Co., US 4 124 748, 1976 (M. Fujimoto, K. Tsuji, T. Nagase).
46. Grain Processing Corp., US 3 661 815, 1972 (T. Smith).
47. Stockhausen Chem. Fab. GmbH, EP 0 068 189, 1983 (M. Chmelir, K. Dahmen, W. Turk).
48. Stockhausen Chem. Fab. GmbH, DE 44 12 153, 1995 (K. Dahmen, R. Peppmöller, U. Günther).
49. Procter & Gamble US 4 654 039, 1985 (K. A. Brandt, S. A. Goldman, T. A. Inglin).
50. Dow Chemical Co., WO 93/21237, 1993 (H. Gartner, P. Trijasson, R. Petri).
51. Kao Corp., US 4 806 578, 1986 (T. Kobayashi, T. Ohya).
52. Nippon Catalytic Chem. Ind., EP 0 632 068, 1994 (Y. Irie, N. Harada, K. Nagasuna, K. Hirota, Y. Masuda, H. Nagano, H. Kubo).
53. Stockhausen Chem. Fab. GmbH, DE 195 43 368, 1995 (L. Breitbach, K. Dahmen, J. Houben, E. Küster).
54. Stockhausen Chem. Fab. GmbH, DE 195 43 368, 1995 (J. Houben).
55. R. A. Scott, N. A. Peppas, *AIChE J.* **43** (1977) 135 – 144.
56. E. Trommsdorff, H. Köhle, P. Lagally, *Macromol Chem.* **1** (1948) 169 – 198.
57. R. G. W. Norrish, E. F. Brookman, *Proc. R. Soc. London* **A 171** (1939) 147 – 171.
58. G. V. Schulz, G. Harborth, *Macromol. Chem.* **1** (1947) 106 – 139.
59. E. Bortel, A. Kochanowski, S. Kudla, E. Witek, *J. Macromol. Sci., Pure Appl. Chem.* **35** (1998) 401 – 409.
60. Hoechst Celanese Corp., US 5 145 906, 1992 (R. D. Chambers, H. H. Fowler jr., Y. Fujiura, F. Masuda).
61. Stockhausen Chem. Fab. GmbH, DE 35 44 770, 1985 (M. Chmelir, J. Pauen).
62. American Colloid Co., US 4 654 393, 1985 (M. Mikita, S. Tanioku, T. Touda).

63. American Colloid Co., US 4 820 742, 1989 (W. Alexander, M. Teppo).
64. Nippon Catalytic Chem. Ind., US 4 625 001, 1985 (T. Tsubakimoto, T. Shimomura, Y. Irie).
65. Nippon Catalytic Chem. Ind., EP 0 508 810, 1992 (Y. Irie, T. Hatsuda, K. Yonemura, K. Kimura).
66. Stockhausen Chem. Fab. GmbH, DE 35 44 770, 1985 (M. Chmelir, J. Pauen).
67. Stockhausen Chem. Fab. GmbH, EP 0 827 443, 1998 (D. Albin, J. Schulte, U. Günther).
68. Stockhausen Chem. Fab. GmbH, DE 195 29 348, 1995 (K. Dahmen, R. Peppmöller).
69. BASF AG, DE 198 46 413, 1998 (J. Dentler, N. Herfert, R. Schliwa, U. Stüven, F. Engelhardt).
70. Nippon Catalytic Chem. Ind., US 4 920 202, 1990 (K. Ishizaki, Y. Irie, K. Iwasaki, N. Harada, T. Hatsuda, K. Kimura, T. Shimonura, T. Fujiwara).
71. Dow Chemical Co., EP 0 513 780, 1992 (H. T. Gartner, A. T. Graham).
72. Stockhausen Chem. Fab. GmbH, DE 40 21 847, 1992 (H. Brehm, K. Dahmen, R. Mertens).
73. Hoechst Celanese Corp., EP 0 463 388, 1992 (G. T. Woodrum, T. H. Majette).
74. Kao Corp., US 4 093 776, 1978 (S. Aoki, H. Yamasaki).
75. Sumitomo Seika Chemicals, EP 0 441 507, 1991 (M. Nakamura, T. Yamamoto, H. Tanaka, H. Ozawa, Y. Shimada).
76. Dow Chemical Co., US 4 708 997, 1987 (F. W. Stanley, J. C. Lamphere jr., Y. Chonde).
77. F. L. Buchholz, *TRIP* **2** (1994) no. 8, 277 – 281.
78. Dow Chemical Co., EP 0 516 925, 1992 (F. W. Stanley, J. C. Lamphere).
79. B. Traub, T. Fritzhanns, S. Hafner, H. W. Spiess, *Colloid. Polym. Sci.* **278** (2000) 547 – 552.
80. Nippon Catalytic Chem. Ind., US 4 734 478, 1985 (T. Tsubakimoto, T. Shimomura, Y. Irie, Y. Masuda, K. Kimura, T. Hatsuda).
81. Nippon Catalytic Chem. Ind., US 4 666 983, 1983 (T. Tsubakimoto, T. Shimomura, Y. Irie).
82. Stockhausen Chem. Fab. GmbH, DE 4 020 780, 1990 (K. Dahmen, R. Mertens).
83. Nat. Starch Chem. Corp., US 4 043 952, 1977 (S. H. Ganslaw, H. G. Katz).
84. Nippon Catalytic Chem. Ind., US 4 734 478, 1985 (T. Tsubakimoto, T. Shimomura, Y. Irie, Y. Masuda, K. Kimura, T. Hatsuda).
85. Stockhausen Chem. Fab. GmbH, DE 4 020 780, 1990 (K. Dahmen, R. Mertens).
86. Dow Chemical Co., US 5 629 377, 1994 (J. H. Burgert, S. B. Christensen, H. A. Gartner, F. L. Buchholz, T. C. Johnson, A. T. Graham).
87. EDANA Recommended Test Methods, European Disposables and Nonwovens Association, Brussels, Belgium (1999).
88. J. D. Ferry: *Viscoelastic Properties of Polymers,* 3rd ed., Wiley, New York 1980.
89. K. Almdal, J. Dyre, S. Hvidt, O. Kramer, *Polymer Gels and Networks* **1** (1993) 5 – 17.
90. Y. L. Yin, Sweling Behavior of Polyelectrolyte Gels, Ph. D. Dissertation, Princeton University, 1993.
91. F. L. Buchholz, A. T. Graham: *Modern Superabsorbent Polymer Technology,* Chap. 4.6.2.1., Wiley, Chichester 1998.
92. Arakawa Chem. Ind., US 4 587 308, 1986 (M. Makita, S. Tanioku).
93. Procter & Gamble, EP 0 631 768, 1995 (M. Schmidt, C. Bewick-Sonntag, M. Plischke).
94. Kimberly Clark Co., EP 0 339 461, 1989 (S. R. Kellenberger).
95. Procter & Gamble, EP 0 691 133, 1996 (M. Plischke, M. Schmidt, S. A. Goldman).
96. Procter & Gamble, US 5 562 646, 1996 (S. A. Goldman, H. L. Retzsch, T. L. Mansfield).
97. I. C. McNeill, A. Alston, *Angew. Makromol. Chem.* **261/262** (1998) 157 – 172.
98. F. L. Buchholz, A. T. Graham: *Modern Superabsorbent Polymer Technology,* Chap. 2.5.4.2., Wiley, Chichester 1998.
99. Stockhausen Chem. Fab. GmbH, DE 198 25 486, 1998 (C. Bruhn, E. Herrmann, J. Issberner, D. Kerstin, R. Mertens, G. Werner).
100. FDA, 21 CFR, Part 177 (1999).
101. Lee County Mosquito Control Di., US 4 818 534, 1989 (R. Levy).
102. Lee County Mosquito Control Di., US 4 983 390, 1991 (R. Levy).
103. H. Sawano et al., *International Wire & Cable Symposium Proceedings* **40** (1991) 333 – 338.
104. M. Fukuma et al., *International Wire & Cable Symposium Proceedings* **36** (1987) 350 – 355.
105. Pirelli Cable Corp., US 4 703 132, 1987 (F. Marciano-Agostinelli, M. Barabaro-Forleo).
106. DSU GmbH, DE 195 29 217, 1996 (H. Kokkelink).
107. DE 196 17 535, 1996 (P. H. Deil).
108. DSU GmbH, DE 195 36 374, 1996 (H. Kokkelink, M. Render).
109. S. Martini, *Melliand* **10** (1998) 717 – 718.
110. Allied Colloids Ltd., EP 0 269 393, 1998 (A. S. Allen, D. Farrar, P. Flesher).

111. Acro Chem. Tech., US 5 026 784, 1991 (B. Le-Khac).
112. Anseal-F product literature, Tyobo Co., Ltd. (1996).
113. O. Grabant, S. Willemsen, J. H. v. Leeuwen, *Proceedings of the International Wire & Cable Symposium* **41** (1992) 293 – 297.
114. Stockhausen Chem. Fab. GmbH, WO 97/00280, 1996 (J. Houben, W. Krug).
115. Procter & Gamble, US 5 338 766, 1994 (D. V. Phan, P. D. Trokhan).
116. BASF AG, WO 97/17397, 1997 (H.-J. Haehnle, M. Walter, J. Tropsch, G. Schornick, T. Anstock).
117. Nova Sorb Ltd. Novel Absorbents, US 5 859 077, 1999 (E. Reichman, A. Skibinsky, D. Kumin).
118. Dow Chemical Co., WO 88/09801, 1988 (R. E. Erickson).
119. Woodbridge Foam Corp., WO 96/31555, 1996 (N. R. Wilson).
120. Stockhausen Chem. Fab. GmbH, DE 44 18 319, 1995 (K. Dahmen, H. Brüggemann).
121. Stockhausen Chem. Fab. GmbH, DE 19 505 708, 1995 (H. Brüggemann, K. Dahmen, D. Lehwald, R. Theilmann).
122. Procter & Gamble, US 5 147 345, 1992 (G. A. Young, G. D. Lavon, G. W. Taylor).
123. Procter & Gamble, WO 93/04113, 1993 (T. A. DesMarais).
124. Procter & Gamble, US 5 571 849, 1996 (T. A. DesMarais).
125. Safety Data of Favor SAB & Favor SX, Stockhausen GmbH & Co. KG, Laboratory for Toxicology and Ecology, Bäkerpfad 25, D-47805 Krefeld.
126. Environmental Safety Data of Favor SAB & Favor SX, Stockhausen GmbH & Co. KG, Laboratory for Toxicology and Ecology, Bäkerpfad 25, D-47805 Krefeld.
127. J. Haselbach, S. Hey, T. Berner, *Regulatory Toxicology and Pharmacology* **32** (2000) 310 – 316.
128. J. Haselbach, S. Hey, C. Jones, *Regulatory Toxicology and Pharmacology* **32** (2000) 326 – 331.
129. J. Haselbach, S. Hey, B. Diener, N. Mehta, *Regulatory Toxicology and Pharmacology* **32** (2000) 317 – 325.
130. H. Greim: *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begründung von MAK- Werten,* 29. Lieferung, Wiley-VCH, Weinheim 1999.
131. S. S. Cutie, S. J. Gonsior in J. D. Hamilton, R. Sutcliffe (eds.): *An Ecological Assessment of Polymers,* Van Nordstrand Reinhold, New York 1996, pp. 103 – 112.
132. G. R. J. Sutherland, J. Haselbach, S. D. Aust, *ESPR – Environ. Sci. & Pollut. Res.* **4** (1997) 16 – 20.
133. J. D. Stahl, M. D. Cameron, J. Haselbach, S. D. Aust, *ESPR – Environ. Sci. & Pollut. Res.* **4** (1997) 16 – 20.
134. *Nonwovens Industry* **27** (1996) no. 11, 40.

Superalloys → Cobalt and Cobalt Compounds
Superalloys → High-Temperature Materials
Superalloys → Nickel Alloys

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Ion Beam Techniques

Standard Article

Gerhard K. Wolf[1] and Wolfgang Ensinger[2]

[1]Universität Heidelberg, Physikalisch Chemisches Institut, Heidelberg, Germany

[2]Universität Marburg, Analytische und Kernchemie, Marburg, Germany

Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.g14_g01
Article Online Posting Date: March 15, 2003

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Interaction of Ions with Matter
3. Equipment for Ion Beam Generation and Application
3.1. Gaseous Ion Sources
3.2. Solid Sources
3.3. IBAD Facilities
4. Applications
4.1. Semiconductor Processing
4.2. Metal Modification
4.3. Ion Beam Modification of Insulators
4.3.1. Ion Implantation and Ion Bombardment
4.3.2. Metallization and Ion Beam Assisted Coatings
5. Outlook

[Top of Page]

1. Introduction

Energetic ions and ion beams are very flexible tools being used in material research, biology and medicine, microelectronics and micromechanics, optics, analysis, and basic research. For most of these fields they provide a certain breadth of applications. In some areas like microelectronics they also play the role of a key technology. The connection and network formation of the ion technology with different areas of science, applications, and materials is shown schematically in Figure 1. Special advantages of ion beams are the flexibility of size ranging from focused beams (≤ 1µm to broad beams ≈ 1 m wide. The penetration depth of the ions in the material ranges roughly from 0.3 nm to 10 µm and depends on their energy. Finally, the beam can consist of ions of nearly any element from hydrogen to bismuth. Because of the possible energy variation from several electronvolts up to 10^9 eV, the differences in mass, and the specific chemical properties of the ions, a huge number of variations exist to generate physical, chemical, or biological effects or analytically useful signals. Ion beams represent an excellent and flexible tool for surface modification of small samples or for basic and applied science applications. In industry they are used for semiconductor processing and for niche applications in different materials finishing procedures. Treatment of large areas is under development but not yet ready for industrial use. Finally, ion beams may be used for analysis of thin films and surfaces (⟶ Surface and Thin-Film Analysis – Static Secondary Ion Mass Spectrometry (SSIMS)).

Figure 1. Connection and network formation of ion beam technology with different areas of science, materials, and application

In the following at first the basics of the interaction of ions with solid materials and the principles of the different ion beam techniques will be described (Chap. Interaction of Ions with Matter). Then a short survey on the necessary equipment and tools will be presented (Chap. Equipment for Ion Beam Generation and Application). In Section Semiconductor Processing . the applications of ion beams for the processing of semiconductors are briefly reviewed. More details are given in Semiconductors – Wafer Fabrication.

Sections Metal Modification and Ion Beam Modification of Insulators deal with the application of ion beam techniques in the modification of metals and insulators, respectively, with special emphasis on features such as wear, friction, corrosion, adhesion, conductivity, and the corresponding microstructural changes. In Chapter Outlook some ideas about future trends will be presented.

For a general introduction into the ion beam technique see [1], [2], [3], [4], [5], [6] and the

- *Proceedings of the biennial International Conference on Ion Beam Modification of Materials IBMM*, published in Nucl. Instr. Meth. B (Semiconductors, Insulators, Metals).
- *Proceedings of the biennial International Conference on Ion Implantation Technology IIT*, mainly published in Nucl. Instr. Meth. B (Semiconductors).
- *Proceedings of the biennial International Conference on Surface Modification of Metals by Ion Beam SMMIB*, published in Mat. Sci. Eng. or Surf. Coat Technol. (Metals, Insulators).
- *Proceedings of the biennial International Conference on Radiation Effects in Insulators REI*, mainly published in Nucl. Instr. Meth. B (Insulators).

[Top of Page]

2. Interaction of Ions with Matter

An energetic ion beam impinging on a solid surface may lose its energy by a number of processes [1], [7]:

1. Electronic excitation of the target atoms (so called electronic stopping).
2. Elastic collisions with the nuclei or atoms of the targets (so called nuclear stopping).
3. Charge exchange and inelastic collisions.

In the energy regime relevant for the applications of ion beam techniques the processes (1) and (2) are of major importance. In the *electronic stopping regime*, the ions (low mass and higher energy) travel straight through the bombarded material, in the nuclear regime (lower energy, higher mass) the collisions lead to strong directional spread.

To give an example: An (α-particle (He ion) of 1 MeV energy loses ≈ 99.9 % of its energy by electronic stopping, an Xe ion of 100 keV energy ≈ 90 % by nuclear stopping.

The straggling of the ions during the slowing down process is responsible for a theoretically Gaussian distribution of the stopped ions in the bombarded material. The half width of the distribution as well as the range depend on the mass and energy of the ions and on the target material. The range is usually expressed as projected range defined as the position of the maximum of the Gaussian curve with respect to the target surface. Referring to the examples given, the projected range of 1 MeV He^+ in Al is 2.6 ± 0.12 µm and of 100 keV Xe^+ only 40 ± 11 nm. The detailed values for most ions and materials may be found in tables or can be calculated by the TRIM code (a computer code based on the LSS theory performed by Lindhard, Scharff, and Schiøtt [10]) [8].

The two dominant stopping modes may cause a number of modifications in the bombarded materials.

- Chemical changes such as cleavage or formation of bonds in compounds.
- Heating of the target.
- Doping or alloying with the implanted ions.
- Generation of various types of defects.
- Enhanced diffusion (surface and bulk).
- Densification of the bombarded region.
- Phase changes and formation of special textures and surface morphologies.

Of special importance is the *sputtering effect*, namely the emission of surface atoms from the target as a consequence of the nuclear stopping process near the surface [2]. Therefore, the sputtering rate is high for low to medium energy of the beam, high ion mass, and low target binding energy. In some cases one impinging ion may sputter 10 – 100 target atoms. Consequently the sputtering process can be used for surface ablation and surface cleaning.

The intensity of the interactions naturally increases with increasing ion dose or ion fluence, expressed as ions/cm^2. Above 10^{12} ions/cm^2 chemical changes and defect formation near the surface of the target can be effected. At fluences around 10^{15} ions/cm^2, modifications of the crystal structure, surface morphology, and electrical properties takes place, and above 10^{15} ions/cm^2 the mechanical surface properties are affected, substantial surface ablation and degradation occur, and the doping and compound formation level approaches the percent regime [3], [9].

As compared with the impact of electrons and photons, ions show the following differences and advantages.

1. Ions interact much more strongly with the target than electrons and photons.
2. The interaction depth of ions is restricted to a few micrometers, however the limits of the interaction zone are sharp and very well defined. There is nearly no depth straggling or interaction zone broadening as in the case of electrons.
3. Ions can be focused to spots < 1 µm allowing direct nanostructuring of surfaces.
4. Ion beams are universal doping tools because any stable ion can form a beam and the energy may be varied from several electronvolts to several megaelectronvolts.

Figure 2. maps schematically the most important basic applications of ion beams. The analytical applications mentioned in the figure are treated in a different keyword (⟶ Surface and Thin-Film Analysis – Static Secondary Ion Mass Spectrometry (SSIMS)).

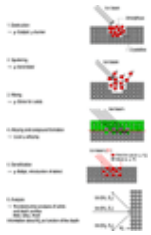

Figure 2. Schematic of the most important properties of the “tool” ion beam

Material modification with ion beams may either be performed by direct bombardment of the subsurface region or by creating a new surface under bombardment by sputtering or by ion assisted deposition. The different techniques are illustrated in Figure 3. [10]:

1. Ion implantation for modifying or doping a region of typically ≤ 1 µm below the surface.
2. Ion beam mixing for improving the interface between an already existing coating and the sample by intermixing and densification (usually restricted to layers ≤ 1 µm thick). For these two techniques an ion implanter or ion source is required that delivers ions of 100 keV – 1 MeV.
3. Ion beam assisted (enhanced) deposition (IBAD). An element or compound is deposited from the gas phase and simultaneously bombarded with an ion beam. This improves the adhesion to the sublayer and densifies or modifies the coating. By choosing the right deposition parameters the structure and texture of the growing film can be affected in a controlled way as compared with pure evaporation. If an element is deposited and the ion beam consists of reactive species or if a reactive gas is fed to the surface during bombardment the coating may be transformed into a compound (Cr + N^+ beam ⟶ Cr_XN, Ti + N_2 + Ar^+ beam ⟶ TiN; see also Nitrides – Boron Nitride).

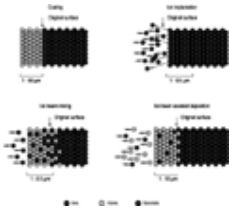

Figure 3. Different ion beam techniques: ion implantation, ion beam mixing and ion beam assisted deposition compared with pure evaporation

For IBAD ion sources are required, that generate low- to medium-energy ions, and an apparatus for evaporating elements and compounds. Some possible combinations are presented in Figure 4.

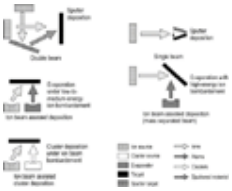

Figure 4. Possible combinations of depositing a coating under ion bombardment (IBAD)

In the case of a focused ion beam or by using masks, structuring or local treatment of a surface is possible as shown in Figure 5. Ions of higher energy may also penetrate deeper into a material, as illustrated in the lower part of Figure 5. This effect can be used for destruction of degenerated tissue for medical purposes or formation of deep implanted layers for microelectronics applications.

Figure 5. Use of focussed and high energy ion beams for A) microstructuring (µ-focussed treatment and alloying < 1 µm); B) formation of deep implants; C) destruction of local regions of tissue

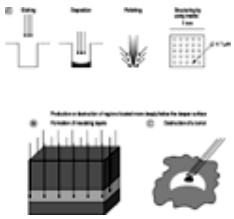

Techniques, which are comparable to ion implantation and IBAD, are plasma immersion ion implantation (PIII) and ion plating [11]. In the case of *plasma immersion implantation* the sample is not bombarded by an ion beam but by ions generated in a plasma. This is performed by giving the sample a pulsed negative bias of several 10 kV. Plasma immersion implantation is less flexible than ion implantation, however, three-dimensional or shaped samples are easier to treat. The advantage of IBAD over ion plating and comparable physical vapor deposition (PVD) techniques (see Thin Films – Physical Vapor Deposition (PVD) Processes) is its lower working pressure and the precise control of the energy, direction of incidence of the ions, and within certain limits also the treatment temperature. However, a disadvantage is the difficulty to treat bigger and odd-shaped items.

[Top of Page]

3. Equipment for Ion Beam Generation and Application

Equipment for ion beam generation and application usually consists of a vacuum chamber, where the beam is formed, and a target station. The probably most important part of an apparatus for ion beam modification is the ion source [4, 5, 12]. An ion source is an ion reservoir from which ions are extracted and form a directed ion beam.

3.1. Gaseous Ion Sources

Here, the reservoir usually is a plasma which is confined in a certain volume. The plasma is sustained by ionized gaseous species, preferably the ones which form the ion beam. However, there may also be an auxiliary gas forming the main part of the plasma while the species of interest is only diluted in it. The most important part of a gaseous ion source is the plasma generator which is basically a vacuum vessel filled with a gas to a certain pressure at a controlled flow rate by means of a valve from an external gas reservoir. Figure 6. shows a schematic presentation of a typical ion source.

In order to ionize the gas atoms or molecules, energy has to be supplied to the gas. A number of processes can be used for ionization. The most common one is electron impact. Electrons are accelerated in an electrical field to kinetic energies high enough to overcome the first ionization potential of the gas particles when they encounter the particles. The electrons can be supplied thermionically by a hot filament. A wire, often made of tungsten or tantalum, is electrically heated to a temperature high enough to emit electrons. In an electrical circuit, this filament acts as cathode, while another part of the ion source, often the vessel itself, acts as anode. Thus, an electric field is established in which the electrons are accelerated. This is the principle of *hot cathode ion sources*.

In *cold cathode ion sources*, ions impinging on the cathode create secondary electrons which in turn are accelerated in the electrical field and ionize the gas. Here, an externally induced ionizing process has to take place in order to ignite the discharge. Often, cosmic rays are sufficient for this purpose.

Mostly, the plasma is confined in a magnetic field. The field elongates the path of the ionizing electrons by forcing them into a spiral movement, thus increasing the number of collisions with gas particles and hence the number of ions.

Apart from electron impact, electromagnetic waves, e.g., *radiofrequency* (rf) waves can be used for ionization. The rf power is coupled inductively or capacitively into the gas volume. The radiofrequency energy excites the particles which become ionized in collisions. In *microwave* ion sources, another wavelength regime often used, electrons absorb the microwave energy and become accelerated to energies high enough for ionization. Further methods of ionization are surface ionization, field ionization, photoionization, spark ionization, etc.

Once ions have been created, they have to be formed to a directed beam. For this purpose, an extraction system is used. An external electrode near the plasma boundary is negatively biased with respect to the plasma and accelerates ions out of the plasma. The extraction electrode is supposed to extract the ions from the plasma, but not to collect them. For collection of ions it contains an aperture. The ions move towards the electrode and through the aperture and then drift on. Thus, an ion beam is formed.

For generating ions of gaseous species in the described way, a number of different ion sources are available. Apart from the kind of ionization, there are other types of classifications which can be applied to ion sources. One is to distinguish sources for narrow but dense ion beams from sources for broad beams, another one makes a distinction between ion sources designed for accelerating the ions to high energies and those which emit low energy ions.

The most prominent source for broad beams of low energy ions is the Kaufman type ion source (see Fig. 6). It is a hot-filament ion source with a large-volume plasma chamber where the ions are extracted by a multi-aperture system. Rather than using an extraction electrode with a single drill-hole, as described below, a large-area grid is used. In this way, an ion beam with a diameter up to several tens of a centimeter can be generated. Another type of ion source capable of creating large area ion beams is the above mentioned radiofrequency-driven source. In contrast to the hot cathode ion sources, which contain one or several filaments made of an oxidizable metal, an rf driven source may be operated with reactive gases such as oxygen or hydrocarbons.

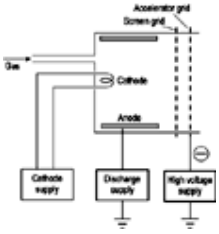

Figure 6. Schematic cross-section of an ion source (hot cathode type for gaseous ions)

Most types of large-area ion sources require a multi-aperture (grid extraction) system. In the end-Hall ion sources, the expensive extraction grids are replaced by a potential difference generated by a magnetic field and a strong electron current.

As mentioned before, the beam shape is determined by the extraction system. Most broad-beam ion sources and focused beam sources, both for low and high energy ions, are equipped with a circular extraction system. These sources emit a circular symmetric ion beam with either a bell-shaped intensity distribution in cross-section of the beam, or with a beam with a more or less flat plateau in the center. When flat elongated workpieces or coil material are to be processed, elongated extraction systems for line-shaped ion beams up to several 10 cm in length are available. In this case, the workpiece is moved through the ion beam. These systems, also called sheet or ribbon beam ion sources, may be scaled-up to a length of 1 m or more.

Broad beam ion sources are used when the beam is directly applied to the workpiece to be modified. However, when the beam has to be guided, swept, or manipulated, its diameter is restricted. In this case, different ion sources with narrow extraction systems have to be used. These sources include the Penning ion gun (PIG) which is of the cold cathode type and therefore capable of generating reactive ions, the Duoplasmatron ion source of the hot cathode type with a particularly intense ion beam with a small diameter (⟶ Surface and Thin-Film Analysis – Ion Sources), the hot-cathode CHORDIS with a comparatively large volume and powerful beam, and the Freeman ion source. These ion sources have typically a single aperture extractor which may be round or elongated. The Freeman source is usually equipped with a slot-shaped extractor.

3.2. Solid Sources

With gaseous feeds, ions of solid materials can only be generated by a molecular species, e.g., B_2H_6, when boron ions are required, or CO for generating carbon ions.

However, in some of the above mentioned ion sources also solid feed materials can be used, namely in sputtering sources and in oven sources. A typical example for the latter is the Freeman ion source in the oven version. In the plasma vessel an electrically heated crucible is placed, out of which solid material with a sufficiently high vapor pressure can be vaporized. Metals such as silver or copper are thermally evaporated into a plasma of an auxiliary gas, mostly argon. Part of the evaporated metal atoms are ionized and can be extracted together with ions of the gas. Another possibility is to use an additional cathode made of the solid material to be ionized. Ions of the auxiliary gas are accelerated towards the cathode, impinge onto it and sputter atoms into the plasma. There, again, the metal atoms are ionized and can be extracted. In this way, ions of metals with a low vapor pressure, such as refractory metals like Cr or Ta, can be generated. Sputtering devices are available in a number of source types, including Duoplasmatron and Penning ion source.

To form an ion beam from a solid without auxiliary gas plasma, metal vapor vacuum arc (MEVVA) ion sources may be applied. Figure 7. shows a schematic cross-section. In this source, an electrical arc is drawn to a cathode. The cathode material is vaporized and forms a high-intensity plasma. These ion sources may also be used for generating broad beams when a broad beam extractor is used. MEVVA ion sources are usually operated in a pulsed mode, in contrast to other ion sources which mostly work in the d.c. mode.

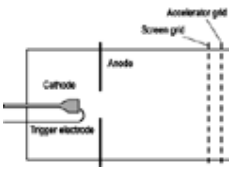

Figure 7. Schematic cross-section of a metal vapor vacuum arc ions source for metal ions

A problem with ion sources operating with auxiliary plasmas is that, as mentioned before, a large number of ions of the auxiliary gas are extracted together with the desired ions of the solid material. Also a certain amount of ions from the ion source construction material may be present as impurity. For this reason, the ion beam has to be purified [6]. This is mostly achieved by means of a sector magnet working as a mass analyzer. When ions with a defined kinetic energy travel through a magnetic field, they are separated according to their mass differences. Light ions are strongly deflected in the magnetic field, while heavy ions follow a smaller deflection radius. By choosing a suitable magnetic field certain ionic species can be selected and pass the magnet, while others impinge on metal plates inside the tube. Another possibility for mass separation is to use a velocity filter based on the Wien principle.

Consequently, ion implantation facilities can be divided in those with mass separation and those without mass separation. Figure 8. shows a schematic representation of a set-up without mass separation. Ions are extracted from a large-area ion source and are accelerated towards the workpiece. For a uniform treatment, the workpiece usually has to be manipulated in the ion beam, e.g., by *x-y* rotation or motion. In case of an implantation facility with mass separation, an ion beam-line

has to be established (see Fig. 9). The ion beam is extracted from the ion source and passes the magnet. After separation, the desired ionic species is accelerated to high kinetic energies by applying a high voltage between the ion source/separation magnet and ground. To this end the ion source is kept at a positive high voltage with respect to the workpiece which is at ground level. Another possibility is to first accelerate the ions and then carry out the separation. In both cases, a beam of highly energetic ions — with kinetic energies from typically several 10 keV up to several megaelectronvolts, depending on the application — is generated. Usually, the ion beam-line contains beam-focusing and deflecting elements such as quadrupoles, Einzel lenses, and electrostatic deflection plates. The ion beam can be focused to a thin intense beam or defocused to a beam with a larger diameter. For a uniform treatment, the ion beam may be electromagnetically or electrostatically scanned in *x*- and *y*-direction and/or the workpiece is moved.

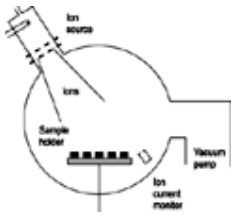

Figure 8. Schematic presentation of broad-beam ion implantation facility (without mass separation)

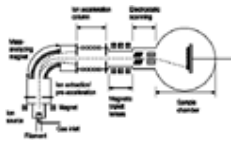

Figure 9. Ion implantation facility with a beam line with mass-separation and ion beam guiding elements

The implantation dose is usually determined by means of measuring the ion current integrated over the process time. Ion beam charge collection is done by a Faraday cup or directly on the conducting workpiece with suppression of secondary electrons.

The workpiece holder is usually a metal plate which can be cooled by water or heated resistively or by IR radiation. Often, it is mounted on a moving facility which allows *x/y*-motion or rotation by means of electromotors. For laterally defined ion irradiation, usually a beam collimator plate with a round or rectangular aperture is mounted in front of it.

All the vessels and tubes are made of vacuum-compatible material, usually with stainless steel seal. For generating the vacuum required for ion beam transport, different types of vacuum pumps such as turbomolecular, oil diffusion or cryopumps, combined with forepumps such as rotary pumps are used. The residual gas pressure is measured by means of Pirani, Penning, membrane, or other vacuum gauges.

3.3. IBAD Facilities

A facility for ion beam assisted deposition is in principle a set-up for broad beam ion implantation, combined with a vapor deposition device (see Fig. 10). Mostly, electron beam evaporation is used because of the versatility and high power input. Electron beam heating allows for vaporizing materials with high melting point and low vapor pressure, both elements such as titanium or carbon, and compounds such as oxides or nitrides. In analogy to the above described sputter ion sources, vaporization can also be achieved by using a secondary ion source which sputters material onto the workpiece (see Fig. 11).

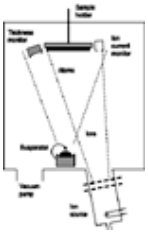

Figure 10. Cross-section of an apparatus for ion beam assisted deposition with electron beam evaporator and ion source

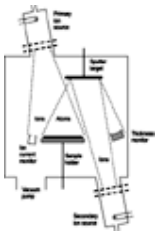

Figure 11. Apparatus for ion beam assisted deposition with one ion source for sputter deposition and another one for ion irradiation of the deposit

Apart from electron and ion irradiation, the materials can also be evaporated electrically by resistive or inductive heating, or by irradiation with intense light (laser).

In IBAD facilities, in addition to gas pressure and ion flux measurement, the deposition rate has to be determined. This is

usually achieved by depositing the film onto microbalances of the quartz crystal monitor type, mounted close to the workpiece, or by measuring optically the atomic vapor density over the crucible. Another possibility is to monitor a property of the deposited film related to its thickness such as its resistivity. Ion and atom measurements define the ion to atom arrival ratio (I/A), an important quantity for IBAD deposition.

In addition to simple IBAD facilities with one ion and one atom source, more sophisticated ones exist with several sources for vapor and ions, these are used particularly when multi-component multi-layer films have to be deposited.

The process chambers for workpiece handling range from volumes of several 10 liters up to cubic meters. The corresponding workpiece holders have sizes up to 1 m^2 capable of handling workpieces up to several 10 kg load.

Facilities for ion beam treatment require specific safety considerations with respect to high voltage and electron induced X-rays. Usually, interlocks and shieldings are used for preventing these hazards. Additionally, toxic feed materials for ion source or evaporator may cause problems. Appropriate gas handling, exhaust and cleaning procedures and devices have to be applied.

For details on ion sources and ion implantation facilities see [4-12, 6].

[Top of Page]

4. Applications

4.1. Semiconductor Processing (⟶ Semiconductors – Wafer Fabrication)

Semiconductor processing is the by far most widespread and important industrial application of ion beam technology. For fabrication of electronic elements the semiconducting material has to be doped, i.e., a small amount of an impurity has to be introduced. Two methods for doping are used, one being diffusion where the dopant is introduced into the semiconductor by thermal motion, the other one being ion implantation. The latter has become more and more important and now plays a greater role. The unique features of ion implantation such as a very good reproducibility and versatility and the possibility to circumvent to a certain extent the restrictions of thermodynamics in order to create uniformly doped regions renders it an excellent tool for well defined doping. Nowadays, worldwide several thousand ion implanters are being used for semiconductor processing. Without ion implantation, the present state of semiconductor technology and the extremely widespread application of semiconductors would not have been reached.

By combination of different technological steps, with ion implantation being the doping step, microelectronic integrated circuit chip technology (IC) is realized. IC manufacturing includes wafer fabrication, wafer probing and sorting, chip assembly, chip testing and burn-in. Both passive electronic elements, namely resistors and capacitors and active circuit devices such as transistors are fabricated in large numbers on a wafer by doping of selected areas by means of ion implantation. The selected areas are produced in a multi-step process by lithography thus creating a desired pattern. The parts which are left free by the mask are doped. The ion beam is directed towards the surface and forms the electronic element in the desired shape and region of the wafer.

The development of IC technology included bipolar transistors and metal-oxide-silicon (MOS) transistors, later, for more power efficiency, complementary MOS technology (CMOS) was established. With this technology, static or dynamic random access memories (SRAMs or DRAMs) are fabricated.

Ion implantation is used for standard CMOS applications, including threshold voltage adjust implants, *n*- and *p*-well implants, field isolation implants, source/drain implants, lightly doped drain implants, punchthrough stop implants, polysilicon doping and others. Advanced applications include ultrashallow source/drain junction formation, channel and drain optimization, trench capacitors and other three-dimensional structures. Figure 12 illustrates the implanted regions and their functions.

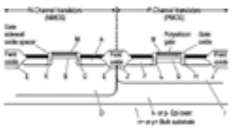

Figure 12. Standard ion implantation applications for CMOS technology

Implanted Regions nd their Functions:

A = NMOS source/drain; basic transistor structure.
B = NMOS channel threshold voltage adjust; sets n-channel V_t (or V_p).
C = NMOS LDD; hot carrier suppression.
D = p-well (“tub”) structure; contains NMOS transistors.
E = p-type “channel stop” for p-well; intra-well (E) and inter-well (E') field isolation.
F = PMOS source/drain; basic transistor structure.
G = PMOS buried-channel threshold voltage adjust; sets p-channel V_t (or V_p).
H = PMOS “punchthrough” suppression.
I = n-well (“tub”) structure; contains PMOS transistors.
J = n-type “channel stop” for n-well; intra-well (J) and inter-well (J') field isolation.
K = NMOS “punchthrough” suppression.
L = PMOS LDD; hot carrier suppresion.
M = polysilicon gate doping (typically n+); improves conductance.

Another ion beam process for semiconductor fabrication is the silicon-on-insulator (SOI) technology. By means of ion implantation at large dose, it is possible to create a silicon layer on top of an insulating silicon oxide layer. Another possibility is a combination of ion implantation and wafer bonding. The surface of a wafer is oxidized. Then hydrogen is implanted through the oxide and a defect layer is formed. Another wafer is bonded to the implanted one. The latter is cut along the defect plane (smart cut process) by means of a heat treatment, forming a new surface of silicon on silicon oxide which is annealed and polished. Thus, a monocrystalline silicon layer on an insulator is fabricated. The SOI material is

dedicated to low-voltage, low-power, high speed systems.

Detailed information on such applications for semiconductors may be found in many monographs [13], [14], articles [16] and the proceedings of the international conferences on, e.g., Ion Beam Modification of Materials (IBMM) and Ion Implantation Technology (IIT).

4.2. Metal Modification

Ion beam modification of metals began in the mid-1970s, mainly for improving wear and corrosion performance. Advantages were seen in the low temperature and the nonequilibrium nature of the process. The conventional metallurgical solubility limits could be exceeded with the possibility to create materials with improved properties.

In contrast to semiconductor processing, ion beam modification of metals has so far only found limited industrial application. The main obstacle for this is the high cost of ion beam treatment. In contrast to most applications in semiconductor processing, metal treatment requires high implantation doses. Additionally, ion beam treatment is in competition with other techniques such as galvanic deposition, and chemical or physical vapor deposition, which yield larger depths of modification or coating. In the following, a number of examples for modification of metals with ion beams will be given.

Nitriding. The certainly most widely investigated application of ion beam treatment of metals is nitriding by nitrogen ion implantation for increased wear life [17-22] (⟶ Metals, Surface Treatment – Nitriding and Nitrocarburizing Methods). The implantation doses typically range from 1 to 8×10^{17} N ions/cm^2, the ion energies from 50 to 100 keV. In general, nitrogen ion implantation leads to hardening. This may be due to different effects, the most important ones being pinning of dislocations, and precipitation of hard phases. In general, nitrides are hard and increase the hardness of the metal in the near-surface zone which, in turn, reduces wear. With ion implantation, nitriding may be performed at room temperature which allows to treat temperature-sensitive metals. Typically, the temperatures can be held below 200 °C. At such temperatures, no phase changes or segregations occur. No distortion of the workpiece is expected, neither does it change its dimensions. Additionally, the surface finish is almost unchanged. Since no coating is applied, there is no adhesive failure to be expected, even at high loads.

Nitriding has been carried out with a large number of alloys, including steels such as austenitic stainless steels and high speed steels, aluminum alloys, titanium alloys such as Ti6Al4V, alloys of other transition metals, including chromium, molybdenum, and tantalum, and hard metals such as tungsten carbide. The implanted nitrogen doses ranged from 10^{17} to 10^{18} N ions/cm^2.

Wear tests under different conditions showed that despite the shallowness of the nitrided zone (well below 1 µm) good results were obtained (see Table 1). This lead to several industrial applications, e.g., in the USA in biomedical applications, where endoprostheses were commercially treated [17], [18]; or in Denmark, where tools for, e.g., plastics processing were modified [19], [20], [24]. Reports on application-oriented studies from other countries such as China, Japan, and Spain may be found in [20], [22], [25-27]. Ion beam processing of tools has been focused on precision tools to avoid alterations in the dimensions or shape by coating or by heating; tools with a refined finish, i.e., highly polished tools; tools heat treated at low temperatures; and very expensive tools because the risk of negative effects in processing is small.

Table 1. Applications of ion implantation in order to increase wear life of tools [17-22]

Application	Life-time increase
Tools for piercing, stamping or punching metal	3–5 ×
e.g., manufacturing of beverage cans (punches)	5–10 ×
Transformation of plastics (moulds, nozzles, valves)	3–4 ×
e.g., extrusion dies for polymer fiber	3–4 ×
Blades for cutting food (vegetables, bread, meat)	4–5 ×
Blades for cutting paper, cardboards, rubber, and plastic	3–7 ×
e.g., blades for cutting synthetic fiber	6 ×
Drills for metal and polymers	2–3 ×
e.g., carbide twist drills for printed circuit boards	2–3 ×

So far, the modified zones have been quite shallow. Their thickness has mainly been governed by ballistic effects which determine the travel distance of the ion in the metal. When the sample temperature is increased, either by an external heating source or, more easily, by energy input via an intense ion beam, diffusion adds as a transport mechanism [28]. In this case, even low ion energies may be used. Energies around 1 keV at ion current densities of several miliamperes per square centimeter were applied. At temperatures of 400 °C, the range of the incorporated nitrogen in austenitic stainless steel has been found to increase by an order of magnitude or more, beyond 10 µm, thus yielding a greatly increased wear life. In the process, a metastable expanded austenite phase may form, another mechanism being the precipitation of CrN, depending on the temperature. The hard nitride phase increases wear resistance, however, it deteriorates the corrosion performance. By an appropriate selection of the process parameters, particularly the temperature, an optimum in wear and corrosion performance can be obtained.

Apart from nitrogen, carbon, oxygen, boron, and a number of metals (⟶ Steel – Vacuum Vapor Deposition) have been implanted for wear reduction.

Formation of Metastable Alloys. As mentioned initially, a particularly strong feature of ion implantation is that virtually any element can be incorporated into any solid. This means that apart from nitrogen any other metalloid and any metal can be implanted into any metallic target. In this way, a metal can be alloyed with any other metal even if they are not soluble in each other. Thus, metastable alloys can be formed. An example is the alloying of iron with mercury, lead, or gold.

Amorphization. The formation of metastable phases is not only possible from a chemical, but also from a structural point of view. Metals are well known for their crystallinity. Ion bombardment can be used to induce an instability in the crystalline structure of metals, thus leading to amorphization [29]. Accumulation of radiation damage may force a metal into a structure without long-range order. Glassy metals exhibit interesting properties, e.g., with respect to corrosion or magnetism (⟶ Metallic Glasses). In conventional techniques the glassy state is obtained by transferring the metal into the liquid state and quenching it by rapidly cooling (splat cooling) the melt on a spinning wheel. This yields ribbons or films of glassy metals. The structural disorder in the melt is frozen in to a maximum amount by using quenching rates as high as possible. When the temperature is far below the recrystallization temperature, the amorphous state will be stable due to a kinetic barrier. The interaction of energetic ions with matter does also yield high quenching rates and is therefore suitable for forming glassy metals. In general, ion irradiation needs large enough ions for creating the required radiation damage and low enough temperatures for preventing recrystallization. For fabricating thin films of amorphous metals, different methods can be used. When intermetallics, e.g., Al_2Au, AlMn, AlFe, Al_3Ti, $AuIn_2$, NiAl, NiZr, FeNd, FeCu, Ni_2Al_3, CuTi, or Zr_3Al, are irradiated at low temperatures with heavy rare gas ions, they accumulate radiation damage and eventually become glassy. Another possibility is to implant metalloids such as P or B which stabilize the glassy state. Rather than bombarding alloys with ions, also multilayers can be irradiated.

Films and Coatings by Ion Beam Mixing. In the case of ion implantation, the base material is always a component of the modified zone. If an alloy or compound of different materials is needed as coating, ion beam mixing can be used. The deposited film is modified by ion irradiation, either structurally or in addition chemically. Thus, a metal can be coated with another metal or a nonmetal which is converted by ion implantation into a compound, e.g., by depositing a chromium film on steel, and implanting nitrogen ions into the film, thus creating a chromium nitride coating. Apart from a chemical change, the microstructural changes induced by ion irradiation play an important role. Prior to ion irradiation the features of the coating are defined by the deposition process. Among these are, e.g., the grain size and the crystal orientation, or, as a result of the microstructure, the porosity and the state of stress. These features can be affected by ion irradiation. By a radiation induced lattice rearrangement, the crystal orientation may be altered. Ion irradiation usually induces a preferential crystallographic orientation. By grain boundary motion, often a change in grain size is observed. The stress state can be influenced. Microporosity, often a consequence of open grain boundaries, can be reduced. In this way, properties such as hardness or corrosion resistance can be influenced. Apart from the film itself its interface to the substrate will be modified when ions travel through it. Figure 13. shows a schematic representation. Chemical bonds will be broken and formed. Impurities can be “diluted”, a mixed zone can be created both by ballistic mixing and by radiation enhanced diffusion. Depending on the combination film-substrate-ion a phase can be generated (interphase). Interface stress can be reduced. These effects may influence important properties of the film-substrate system, such as its porosity or its ohmic resistance. The probably most important property is the adhesion between film and substrate. The above mentioned effects usually lead to substantial enhancement of adherence between the film and the metal substrate.

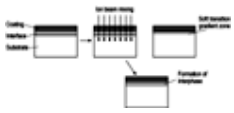

Figure 13. Ion beam mixing of a coating on a substrate, forming a transition zone (gradient) or an interface. Both may enhance adhesion.

As ion beam mixing is essentially also limited by the range of the ions, it suffers from similar restrictions with respect to thickness of the modified zone as ion implantation. For metal modification often thicker modified zones are needed, e.g., for corrosion protection. The problem with shallowness of the modified zone can be overcome when thin films are deposited from the vapor phase under simultaneous rather than postdeposition bombardment with energetic ions (IBAD). Simultaneous ion irradiation or implantation influences the film in a way similar to ion beam mixing. In situ ion bombardment is particularly powerful with respect to film densification, preferential crystallographic orientation, grain size and morphology, and stress [30-32]. Thin films usually suffer from stress, either due to a mismatch in the thermal expansion coefficient or intrinsic stress, which arises from defects, impurities etc. Often, the stress state of physical vapor deposited films is tensile. By means of simultaneous ion irradiation, it may be reduced or changed to compressive stress. Figure 14. shows a typical dependence of the stress state on the I/A-ratio. The stress state changes from tensile to compressive, with a state of zero stress in the transition. This allows to deposit films with low stress levels, thus enhancing adhesion.

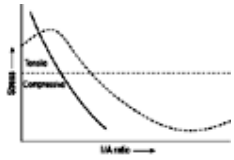

Figure 14. General dependence of stress of thin IBAD films on the ion to atom arrival ratio (I/A)

IBAD coatings range in their thickness from 0.1 µm up to 10 µm with a typical thickness of 1 to 2 µm. The coatings can be divided in elemental or compound layers and in single layers (monolithic) and multilayers. Films of a single element consist mostly of metals such as Al or transition metals like Ti, Cr or Ta, nonmetals such as Si, B, or, mostly, carbon. The latter is

usually an amorphous material with varying ratio in sp^2/sp^3 hybridization, sometimes including hydrogen. Among the compound films, the nitrides play a major role. TiN, ZrN, CrN, TaN and other transition metal nitrides, AlN from the main group metals, and nonmetal nitrides, particularly Si_3N_4, CN_x, and BN, especially the hard metastable cubic modifications, have been deposited on many kinds of metal substrates. All these can comparatively easily be formed by evaporating the element and irradiating the growing film with nitrogen ions. Another group of compounds are the oxides, such as Al_2O_3, SiO_2, TiO_2, or Ta_2O_5. When the element is highly reactive, the film can be formed by nonmetal ion implantation and reactions of the growing film with gas from the atmosphere. In this case, both reactive ions (nitrogen or oxygen) or rare gas ions, mostly argon, can be used. The most important and typical examples of nitride and oxide films on metals formed by reactive ion beam assisted deposition are TiN and Al2O3. When the element is not reactive enough, particularly in case of the main group elements, the nitrogen or oxygen has to be incorporated by ion implantation. Here, the I/A ratio directly determines the composition. Examples are AlN, SiO_2, Si_3N_4 and BN. While the metallically bonded transition metal nitrides can be grown crystalline over a wide temperature range, the covalent oxides and nitrides of the main group elements tend to grow amorphous.

Microlaminates. When different materials are deposited in alternating order or the ion beam is modulated, *multilayers* can be formed. These exhibit often considerably better features then single films of either of the materials, such as enhanced elasticity and hardness. Particularly, their mechanical strength can be greatly improved. Such so-called microlaminates have a great stability both against crack initiation and propagation. A typical examples is the system Al-Al_2O_3. The microlaminate shows a much lower crack initiation and propagation probability than either of the monolithic layers.

Wear protection. Most of the nitrides and oxides mentioned above are harder than metals. Therefore, they may serve as protection against wear. The most prominent of the nitrides is certainly TiN. It may be deposited by evaporation of titanium either under nitrogen ion bombardment or under rare gas, usually Ar, irradiation in an atmosphere of backfilled nitrogen. Depending on the process parameters titanium nitride may have a broad range in composition and also in hardness. Both comparatively soft and ductile films may be obtained with hardness well below that of bulk TiN (20 GN/m^2), but also a quality with considerably larger hardness can be obtained, where the films usually have intrinsic compressive stress. Friction and wear tests of TiN films on different substrates, including stainless steel, aluminum alloys, but also tungsten carbide, gave reductions both in friction coefficient and wear. Figure 15. shows the dependence of hardness and wear life of TiN formed by reactive IBAD on the Ar^+/Ti arrival ratio.

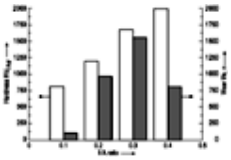

Figure 15. Dependence of hardness (left scale) and wear life (right scale, given as number *n* of passes of the wear test) of TiN formed by reactive IBAD on the Ar^+/Ti arrival ratio [32]

Of particular interest is the group of (potentially) ultrahard materials, namely diamond-like carbon (DCL) and cubic boron nitride (c-BN). Both can be grown by ion beam assisted deposition. Thin DLC films are often comparatively elastic and exhibit particularly low friction coefficients. Cubic boron nitride films show similar features and may improve the tribological performance of metals. However, the process window is comparatively narrow, and the films suffer from large compressive stresses which may cause adhesion problems.

Lubricant Coatings. Apart from hard coatings, solid lubricants can be used. Molybdenum sulfide (MoS_2) grown either as monolithic film or as multilayer in combination with hard coatings such as B, B_4C, or BN, show very low friction coefficients.

Corrosion Resistant Alloys. Metals may easily be alloyed by ion implanting an alloying element which leads to a more corrosion resistant alloy [33], [34]. An example is the implantation of chromium into iron which, if the implantation concentration exceeds 12 at%, becomes corrosion resistant in neutral aqueous media similar to stainless steels. In the literature, a large number of examples of ion implantation into various metals and alloys, including alloys based on iron, aluminum, titanium, nickel, and others are described. The implanted species were all kinds of metals and nonmetals including noble metals such as platinum or palladium, and metalloids such as phosphorus and boron. The layers formed protect against uniform corrosion, localized corrosion such as pitting and crevice corrosion, and hydrogen embrittlement, which may occur when metals are exposed to water, acids, or salt brines. However, the effect is often not long-lasting because the modified zone below the surface is too shallow. When corrosion progresses, the thin modified zone may be consumed in a short time, ranging from a few days to weeks. For particular cases a long-term protection can be achieved. The cathodic protection offered by implanting catalytically active noble metals such as platinum is very effective for some metals with active/passive transition behavior. These metals such as titanium are protected by a native oxide which leads to passive corrosion with low dissolution rates. In aggressive environment, the oxide may be dissolved with the result that active corrosion with high rates starts. The noble metal shifts the potential into a region where the oxide film is recovered and the metal remains passive. When the implanted protective material itself is not consumed by corrosion, it may remain on the metal surface and exert a long-term protection effect. In this way, titanium can be protected against corrosion in mineral acids or salt brines, and tantalum can be protected against corrosion and hydrogen embrittlement in hot concentrated mineral acids.

When the implanted zones are too thin and/or the effect is not long-lasting, ion beam assisted deposition of thicker films may be applied. Usually, a coating more corrosion resistant than the base material is used. It may consist of nonmetals such as boron, silicon, or carbon, resistant metals such as noble metals, but also aluminum, titanium or tantalum, or nitrides, carbides and oxides. It turns out that the most important characteristics of corrosion-resistant thin films deposited under ion bombardment are low microporosity and high adhesion [34]. These features can be optimized by proper selection of the main deposition process parameters which are I/A ratio, ion energy, and ion incidence angle. In the

beginning of film growth, ion beam mixing effects lead to a strengthening of the interface and to enhanced adhesion. During film growth, ion bombardment increases the density of the film material and closes micropores. Often, the closing of pores is enhanced when an off-normal ion incidence angle is chosen.

Again, multilayer films may be more effective than monolithic films. Figure 16. shows the iron dissolution currents of an iron sample coated with a Ti/TiN multilayer sequence. The multilayers are less porous than the corresponding single films of either Ti or TiN. It can be assumed that monolithic films contain pores from the substrate through the film to the surface. TiN grows in columns parallel to the surface normal. Between the columns, more or less open channels may exist. In a multilayer film with a sequence of a cubic nitride and a hexagonal metal crystalline structure, such continuous open structures are less likely to form.

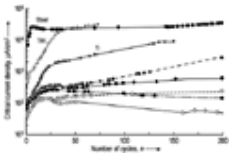

Figure 16. Critical current densities deduced from current-potential plots for Ti/TiN multilayers deposited by IBAD on steel vs numbers of potential cycles

(●) CK45 steel, (*) Ti, (+) TiN, (♦) Ti/TiN ratio: 100/25, (○) ratio: 100/50, (■) ratio: 100/100, (□) ratio: 50/100, (▲) ratio: 25/100 [36]

Not only aqueous corrosion, but also high temperature corrosion in gases can be reduced. Coatings such as Si3N4 are effective in reducing oxidation of metals such as Ni and Cr and metallic alloys.

Films with Special Electric and Magnetic Properties. Apart from mechanical and chemical features, electric and magnetic properties can be influenced. For metal films, a strong dependence of the electric resistivity on ion energy, I/A-ratio, and type of metal has been found [35]. Magnetic properties such as coercivity can be influenced and films with giant magnetoresistance can be formed.

4.3. Ion Beam Modification of Insulators

The effects of ion beams in insulators are somewhat different from those in metals because:

- The electronic stopping of ions causes much more changes in insulators than in metals.
- Insulators are mostly compounds, which means that the incoming ions interact with chemical bonds during slowing down.
- The bombardment of insulators with ions may cause strong charging of the samples, an effect which has to be taken into account during the experiment.

The application of ion beams for insulator modification can be grouped in two categories:

1. Ion implantation and ion bombardment.
2. Metallization and ion beam assisted coatings.

The most important materials treated in practice are different types of polymers, engineering ceramics, and glasses.

4.3.1. *Ion Implantation and Ion Bombardment*

Polymers. Most polymers are not very stable to radiation, which means that already moderate fluences of heavy energetic ions may partly destroy the polymer structure. Figure 17. indicates the chemical changes caused by the ions with increasing dose and the corresponding variations of properties. In the low dose range mainly properties such as solubility or wettability are affected, in the higher dose range — near the region where graphite is formed — also hardness, conductivity, and optical properties are modified.

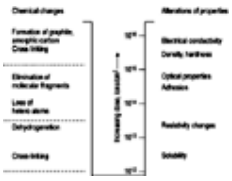

Figure 17. Dose dependence of the chemical and physical properties of polymers and the corresponding structural changes

Generation of Local Conductivity. Many researchers tried to induce conductivity in polymers, because this is an interesting possibility to obtain locally conductive regions in an insulating material, for example for O-electronics applications. Figure 18. illustrates the variations of surface conductivity of the photoresistant polymer HPR-204 as function of the impinging dose of 2 MeV Ar^+ ions. The conductivity increases by four to five orders of magnitude and reaches the values of ion bombarded amorphous carbon or disordered metals [37].

Figure 18. Resistivity changes of 2 MeV Ar^+ ion bombarded polymers (HPR-204) as function of ion dose in comparison with amorphous carbon

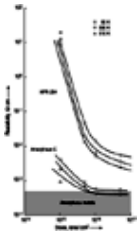

Another possibility for obtaining local conductivity is the irradiation of intrinsically conductive polymers such as polypyrrole or polythiophene. In this case the irradiation destroys the conductive properties, and, if done through a mask, leaves only the shaded areas conductive. Figure 19. shows an example of such a structure with subsequent copper deposition on the conductive areas [38], [39].

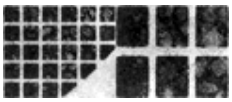

Figure 19. Galvanic deposition of copper on the shaded areas of a conducting polymer bombarded with argon ions through a mask (size of the structures 13.5 µm)

Production of Microfilters. Microfilters can be produced by high energy ion beams (typically 100 MeV). The stopping track of the ions in polymers can be etched by alkaline solutions into very small channels suitable for special filtering or separation problems [40].

Glass. In the case of glass ion beam techniques have mainly been used to change its *optical properties*. These alterations range from the implantation of color centers (rare earth ions, for example) with the aim of flat screen production to the generation of wave guides for light. In the latter field a number of groups are working. The basic principle is the local implantation of metals which are not soluble in the glass and thus form small precipitates, causing local changes in the reflectivity of the materials [41], [42].

In addition ion beams are also used for polishing or machining high precision optical lenses in the nanometer-regime by sputtering.

Ceramics. Here the improvement of the mechanical surface properties like hardness, friction, and cracking behavior were the main subjects of application oriented studies. Especially nitrogen implantation in hard metals and in silicon nitride showed a considerable potential for crack reduction and the formation of intermediate solid lubrication layers [43], [44].

4.3.2. *Metallization and Ion Beam Assisted Coatings*

Ion beam mixing and ion beam assisted deposition are excellent methods for metallizing polymers, ceramics, and glass. In addition also compound layers or multilayers can be deposited on these substrates.

The advantage of ion beam techniques in comparison with pure evaporation is the greatly improved adhesion, and the low processing temperature, in comparison with PVD methods. The latter is especially important for the polymer case.

Metallizing of Polymers. Metallized polymers find broad applications in microelectronics as base materials for circuit boards, for electronic shielding in several industrial branches and for decorative and package purpose. Well established techniques are galvanic deposition and vacuum evaporation with or without prior plasma cleaning and roughening. Because of the low processing temperature and because they are not entailed with surface roughening, ion beam techniques can be used in cases were very smooth surfaces are needed, or for polymers which are difficult to metallize, such as high temperature thermoplastics [45], [46].

The factors affecting adhesion of metal layers on substrates are compiled in Figure 20. Most of these factors can be modified by ion bombardment, especially the film stress and the chemistry and morphology of the interface. This is illustrated in Figure 21. were the adhesion of thin copper layers on polyethersulfone (PES), measured by the peel test, is plotted as a function of the mixing ion species and the ion energy [47]. Before peeling the original Cu-layers have to be thickened to 20 µm by galvanic means.

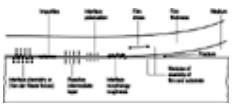

Figure 20. Schematic representation of the factors affecting the adhesion of metal films on polymer or ceramic substrates

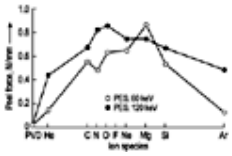

Figure 21. Adhesion of copper deposited on polyethersulfone (PES) after bombardment with different ion species measured by the peel test (ion dose $10^{14}/cm^2$; ion energy 60 keV or 120 keV; PVD: deposition without bombardment)

A layer deposited by vacuum evaporation shows initially no adhesion, upon bombardment the adhesion first increases with increasing electronic energy loss of the ion beam and goes through a maximum. The decrease after the maximum is caused by the increasing importance of nuclear energy loss in the higher mass region, leading to a destruction of the polymer skeleton. Similar effects have also been observed using ions with higher energy in the range of a few 100 keV–10 GeV [48]. The IBAD technique can also be used for improving the metal adhesion on polymers. The obtainable effects are very much dependent on the chemical composition of the polymer. In the case of polyethersulfone only very small ion doses (expressed in terms of the ion to atom arrival ratio I/A) of argon ions in the energy range of 2.5 – 5 keV or higher doses of low energy ions of 0.5 – 1 keV lead to considerable improvements. In Figure 22. the adhesion force, measured by the pull off test, is plotted as function of I/A for several energies. Interestingly, especially polyphenylenesulfide (PPS) is rather difficult to metallize with the established techniques. In some cases also a Ti intermediate layer improves the situation.

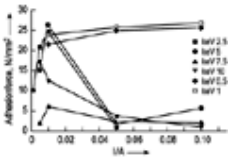

Figure 22. Adhesion force of a copper layer deposited on polyphenylensulfide (PPS) by ion beam assisted deposition as function of the ion/atom arrival ratio, measured by the pull off test (parameter is the argon ion energy)

(♦) 0.5 keV; (|) 1 keV (■); 2.5 keV; (●) 5 keV; (▲) 7.5 keV; (▼) 10 keV

Metallization by IBAD has been used also for medical purposes. Polymer tubes which are applied for “in body” therapy of kidney diseases have been coated with thin silver layers of excellent adhesion. Silver has bactericidal properties and thus helps to avoid unwanted inflammations.

Metallization of Ceramics. Only little work has been done on metallization of ceramics by ion beam mixing or IBAD. One example for applications is again the formation of conductive layers on an insulating substrate, another one are adhesive coatings with low friction coefficient. Si_3N_4 and Y stabilized ZrO_2 (PSZ) were coated with a number of metals and subsequently irradiated with 140 keV Ar^+ ions.

Under the running conditions of a Diesel engine the friction coefficient was reduced by nearly one order of magnitude for the case of a Ni–Ti double layer [49].

[Top of Page]

5. Outlook

At present the cost of non mass analyzed ion implantation (nitrogen for example) is ≈ 0.20 € for an implanted dose of 10^{15} ions/cm^2 and ≈ 0.70 – 1 € for 10^{17} ions/cm^2. Metal ion implantations with the MEVVA source was estimated in the USA to range around 0.10 €/cm^2.

These figures show that modifications requiring low ion doses, like in the case of polymers, are relatively cheap to perform. Also the PIII technique is somewhat less expensive than ion beam implantation. Ion beam assisted deposition is already used routinely in the optical industry and for small series or development work in many other industrial fields. The costs of the modification depend on one hand on the investment, which is usually 10–30 % higher than that for a conventional PVD plant. Secondly the beam power of the ion source and the thickness of the coating determine the cost per square centimeter. A 3–5 µm thick TiN layer may be realized at present for 1–2 €/cm^2. With an optimized dedicated machine 0.25 €/cm^2 should be no problem.

At present plants with continuous coil coating or large area capability using powerful slit extraction sources are under development [50]. With such plants cost reductions may be realized, especially for cases where the ion beam is only used for precleaning by sputtering or for the production of thin interface or barrier coatings. The thickening of the coating is then performed by pure evaporation.

The future trends for applications of ion beam techniques are on the one hand dependent on the further development of high power broad beam sources and continuous working plants, on the other hand on the future requirements for high-tech coatings. Fields of high potential for special surfaces and coatings are:

- Implants and materials for medical and biological use like internal prostheses, screws, fixations, infusion tubes, catheters etc.
- UV- and X-ray optics, specially surface finishing of lenses by sputtering with nanometer precision.
- Texturing and metallizing of polymers and ceramics for circuit boards and sensors.
- Coatings and nitriding of highly temperature sensitive materials like Al, Mg, and their alloys.
- Surface modification and machining of micromechanical devices and creation of nanostructures by using mask techniques or highly focused ion beams.

[Top of Page]

References

General References

1. G. K. Wolf: “Chemical Effects of Ion Bombardment” in: *Topics in Current Chemistry*, vol. **85**. (1979), p. 1.
2. G. Carter, J. S. Colligon: “Ion Bombardment of Solids, Chapter Sputtering”, *Heinemann Educational Books*, London

1968, p. 310.

3. H. Oechsner, J. Walldorf, G. K. Wolf: Teilchenstrahlgestützte Verfahren in G. Kienel, K. Röll (eds.): *Vakuumbeschichtung* **2**, VDI-Verlag Düsseldorf, 1995, p 193.
4. J. J. Cuomo, S. M. Rossnagel, H. R. Kaufman (eds.): *Handbook of Ion Beam Processing Technology*, Noyes, Park Ridge 1989.
5. J. F. Ziegler (ed.): *Handbook of Ion Implantation Technology*, North-Holland, Amsterdam 1992.
6. B. H. Wolf (ed.): *Handbook of Ion Sources*, CRC Press, Boca Raton 1995.

Specific References

7. J. Lindhard, J. M. Scharff, H. E. Schiott, *Mat.-Fys. Medd.-K. Dan. Vidensk. Selsk.* **33**, (1963) 14.
8. J. F. Ziegler, J. P. Biersack, U. Littmark: *The Stopping and Range of Ions in Solids*, Pergamon Press, New York 1985.
9. G. K. Wolf: Ionenstrahltechnologie – Werkzeug für das 21. Jahrhundert, *Vakuum in Forschung und Praxis* **3** (1996) 143.
10. G. K. Wolf: “Erzeugung verschleißmindernder Randschichten durch Ionenstrahlverfahren”, *Materialwis. Werkstofftech*. **24** (1993) 109.
11. J. R. Conrad, “Plasma Source Ion Implantation: A New Approach to Ion Beam Modification of Materials”, *Mater. Sci. Eng.* **A 116** (1989) 197.
12. W. Ensinger, “Ion Sources for Ion Beam Assisted Deposition of Thin Films and Coatings”, *Rev. Sci. Instrum.* **63** (1992) 5217.
13. W. S. Ruska: *Microelectronic Processing: an Introduction onto the Manufacture of Integrated Circuits*, McGraw Hill, New York, 1987.
14. D. J. Elliott: *Integrated Circuit Fabrication Technology*, McGraw Hill, New York, 1989.
15. T. Seidel, “MOS Process Review with Implications for Ion Implantation”, *Nucl. Instr. Methods* **B21** (1987) 96.
16. E. Chason *et al.*, “Ion Beams in Silicon Processing and Characterisation”, *J. Appl. Phys.* **81** (1997) 6513.
17. P. Sioshansi, “Surface Modification of Industrial Components by Ion Implantation”, *Mater. Sci. Eng.* **90** (1987) 373–383.
18. J. K. Hirvonen: “Ion Beam Processing for Industrial Applications”, *Mater. Sci. Eng.* **A116** (1989) 167–175.
19. B. Torp *et al.*: “A New Industrial Ion Implanter for Surface Modification of Metals”, *Nucl. Instr. Meth. Phys. Res.* **B127/128** (1997) 940–944.
20. G. I. Zhang: “Present Status of Ion Implantation in Non-Semiconductor Materials in China”, *Mater. Sci. Eng.* **A115** (1989) 377.
21. M. Geng *et al.*: “Modification Technology of N^+ Implantation of Carbide Twist Drills Used in Integrated Circuit Boards and its Industrial Application”, *Surf. Coat. Technol.* **51** (1992) 551–555.
22. H. Li, C. Ji: “Surface Modification of Non-Semiconductors by Ion Beams in China”, *Surf. Coat. Technol.* **65** (1994) 180–193.
23. C. A. Straede, N. J. Mikkelsen: “Implementation of Ion Implantation in European Industry”, *Surf. Coat. Technol.* **84** (1996) 567–573.
24. C. A. Straede, N. J. Mikkelsen: “Industrial Implementation of Ion Implantation on Tools and Wear Parts Based on a Strategic Approach”, *Surf. Coat. Technol.* **103–104** (1998) 191–194.
25. M. Iwaki: “Ion Implantation in Japan in Non-Semiconductor Fields”, *Mater. Sci. Eng.* **A115** (1989) 369–376.
26. R. J. Rodriguez, A. L. Sanz, A. M. Medrano: “The Search for New Applications of Ion Implantation Treatments”, *Surf. Coat. Technol.* **84** (1996) 594–599.
27. J. I. Onate, F. Alonso, I. Braceras, A. L. Sanz, R. J. Rodriguez: “Current Status of Commercial Ion Implantation in Spain”, *Surf. Coat. Technol.* **103–104** (1998) 185–190.
28. R. Wei: “Low Energy, High Current Density Ion Implantation of Materials at Elevated Temperatures for Tribological Applications”, *Surf. Coat. Technol.* **83** (1996) 218–227.
29. P. M. Ossi: “Ion Induced Crystal-to-Glass Transition in Alloys”, *Mater. Sci. Eng.* **A115** (1989) 107-121.
30. F. A. Smidt, G. K. Hubler: “Recent Advances in Ion Beam Modification of Metals”, *Nucl. Instr. Meth. Phys. Res.* **B80/81** (1993) 207–216.
31. J. K. Hirvonen: “Current Topics of Ion Beam R&D”, *Surf. Coat. Technol.* **65** (1994) 84–89.
32. W. Ensinger: “Ion Bombardment Effects during Deposition of Metal and Nitride Films”, *Surf. Coat. Technol.* **99** (1998) 1.
33. G. K. Wolf: “A Historical Perspective of Ion Bombardment Research for Corrosion Studies”, *Surf. Coat. Technol.* **83** (1996) 1–9.
34. W. Ensinger, “The Influence of Ion Irradiation during Film Growth on the Chemical Stability of Film/Substrate Systems”, *Surf. Coat. Technol.* **80** (1996) 35.
35. R. Roy: “Low-Energy Ion Assisted Deposition of Metal Films”, *Surf. Coat. Technol.* **51** (1992) 203–211.
36. R. Hübler, A. Schröer, W. Ensinger, G. K. Wolf, W. H. Schreiner, I. J. R. Baumvol: “Plasma and Ion – Beam – Assisted Deposition of Multilayers”, *Surface and Coatings Technol.* **60** (1993) 561.
37. T. Venkatesan: “High Energy Ion Beam Modification on Polymer Films”, *Nucl. Instr. Methods* **B. 7/8** (1985) 461.
38. S. Schiestel, P. Banniza, G. K. Wolf, K. Edinger: “Modification of Intrinsically Conducting Polymers by Ion Implantation”, *Nucl. Instr. Methods* **B. 116** (1996) 164.
39. T. Asmus, G. K. Wolf: “Modification and Structuring of Conducting Polymer Films on Insulating Substrates by Ion Beam Treatment”, *Nucl. Instr. Methods* **B 166/167** (2000) 732.
40. R. Sudowe, P. Vater, R. Brandt, J. Vetter, W. Ensinger: “Filters with < 100 nm Radius Pores for Gas Separation

Formed by High Energy Ion Irradiation of Polymers", *Nucl. Instr. Methods* **B 175–177** (2001) 564.

41. F. Gonella: "Nanoparticle Formation in Silicate Glasses by Ion-Beam-Based Methods". *Nucl. Instr. Methods* **B 166/167** (2000) 831.
42. S. Schiestel, S. B. Quadri, C. A. Carosella, G. K. Hubler, D. Knies: "The Control of Gold Nanocluster Sizes and Volume Fraction in Dieletric Thin Films via Ion-Beam Assisted Deposition", *Nucl. Instr. Methods* **B 175–177** (2001) 314.
43. A. Mucha, W. Fischer, G. K. Wolf: "Modifizierung von keramischen Oberflchen", *Ingenieur-Werkstoffe* **5** (1993) 52.
44. W. Fischer, H. Wituschek, G. K. Wolf, H. Ferber, R. Heinze, M. Woydt: "Modification of the Mechanical Surface Properties and Tribo-Chemistry by Ion Beam Techniques", *Surf. Coat. Technol.* **59** (1993) 249.
45. H.-J. Kersten, G. K. Wolf: "Metallization of High-Temperature Thermoplasts by means of Ion-Beam-Assisted Deposition", *Surf. And Coatings Technology* **116–119** (1999) 1183.
46. H. Kupfer, G. K. Wolf: "Plasma and Ion-Beam-Assisted Metallization of Polymers and their Application", *Nucl. Instr. Methods* **166–167** (2000) 722.
47. N. Tegen, G. K. Wolf, H. Wartusch, H. Merkel: "Chemical and Energy Deposition Effects of Cu-Films onto Polymers", *Nucl. Instr. Methods* **B 80/81** (1993) 1055.
48. D. M. Rück: "Ion Induced Modification of Polymers at Energies between 100 keV and 1 GeV Applied for Optical Waveguides and Improved Metal Adhesion", *Nucl. Instr. Methods* **B 166–167** (2000) 602.
49. W. Wei, J. Lankford, R. Kossowski: "Friction and Wear of Ion-Beam-Modified Ceramics by Surface Modification", *Surface and Coatings Technology* **37** (1989) 179.
50. G. K. Wolf, G. Preiß, R. Münz, L. Guzman: "Large Area IBAD Deposition of Zn-Alloys in the Coil Coating Mode", *Nucl. Instr. Methods* **B 175–177** (2001) 756.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Inulin

Standard Article

Anne Franck[1] and Bart Levecke[1]
[1]ORAFTI, Tienen, Belgium

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.g14_g02
Article Online Posting Date: July 31, 2003

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Chemical Structure
3. Physical and Chemical Properties
4. Physiological Properties
5. Resources and Raw Materials
5.1. Dahlia
5.2. Jerusalem Artichoke
5.3. Chicory
6. Production
6.1. First Step
6.2. Second Step
6.3. Environmental Aspects
7. Uses
7.1. Food Applications
7.2. Inulin Derivatives for Non-Food Applications
7.2.1. Oxidation
7.2.2. Carboxymethylation
7.2.3. Hydrophobization
8. Analysis
8.1. HPLC
8.2. Gas Chromatography
8.3. HPAEC Analysis (Dionex)
8.4. Permethylation
8.5. Quantitative Determination
9. Legal Aspects
10. Economic Aspects and Trade Names

[Top of Page]

1. Introduction

Inulin is a storage carbohydrate found in many plants. It has formed part of the daily diet of humans for some centuries as it naturally occurs in many vegetables, fruit, and cereals. Inulin is industrially obtained from chicory roots and used as a food ingredient offering both interesting nutritional properties and important technological benefits. It significantly improves the organoleptic characteristics of food and drinks, increases the stability of foams and emulsions, and shows a fat-like behavior when used as a gel in water.

Inulin was first isolated from *Inula helenium* by ROSE, a German scientist, in 1804 [1] but it was THOMSON who called this substance inulin [2]. One pioneer in fructan research was the German plant physiologist JULIUS SACHS [3] who detected inulin in the tubers of *Dahlia*, Jerusalem artichoke (*Helianthus tuberosus*), and *Inula helenium* after ethanol

precipitation. Already in the beginning of the 20th century, feeding diabetic patients with inulin was reported to be beneficial [4]. Especially since the 1990s, a spectacular increase in the number of publications dealing with the technological and nutritional benefits of inulin has been observed [5].

[Top of Page]

2. Chemical Structure

Inulin [9005-80-5] is a polydisperse carbohydrate material consisting mainly, if not exclusively, of fructose [57-48-7] units, linked by β(2 ⟶1) bonds [6]. A starting glucose [50-99-7] moiety can be present in the chain but is not necessary. *Levan*, formed by certain bacteria, consists mainly or exclusively of β(2 ⟶6) linked fructosyl-fructose units. As for inulin, glucose can be present, but again is not necessary. *Fructan* is a more general name, used for any compound in which fructosyl-fructose bonds constitute the majority of linkages. The term fructan therefore covers both inulin and levan.

Referring to the definition of inulin, both GF_n and F_n compounds (F being a fructosyl unit and G a glucosyl unit) are included under that same nomenclature. In chicory inulin, *n*, the number of fructose units, can vary from 2 to 60 [7], i.e., inulin is a mixture of oligomers and polymers. Its molecular structure is shown in Figure 1.

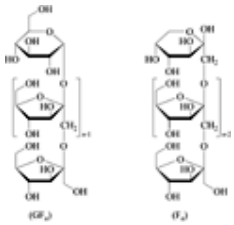

Figure 1. Chemical structure of inulin

The degree of polymerization (DP) and the possible presence of branches are important properties because they influence the functionality of inulin. Therefore a strict distinction must be made between inulin from plant origin and inulin from bacterial origin. The DP of plant inulin is rather low (DP < 200) and varies according to the plant species, weather conditions and the physiological age of the plant. Standard chicory inulin has an average DP of about 10–12. Long-chain chicory inulin, from which the lower DP fraction has been physically removed and which has an average DP of about 25, is also available. In contrast, bacterial inulin has a very high DP, ranging from 10000 to over 100000 (10^6 to 10^7 dalton). It is branched (≥15 %) which results in a completely different behavior in aqueous systems.

[Top of Page]

3. Physical and Chemical Properties

Chicory inulin is available as a white, odorless powder with a high purity and a defined chemical composition. The taste is neutral, without any off-flavor or aftertaste. Standard chicory inulin is slightly sweet (10 % sweetness in comparison with sucrose), whereas long-chain inulin is not sweet at all [5] (see also Physiological Properties). The physicochemical properties of chicory inulin are summarized in Table 1.

Table 1. Physicochemical properties of chicory inulin

	Standard inulin	Long-chain inulin
Chemical structure	GF_n ($2 \le n \le 60$)	GF_n ($10 \le n \le 60$)
Average degree of polymerization	12	25
Dry matter (%)	> 95	> 95
Inulin content (% of d.m.)	92	99.5
Sugar content (% of d.m.)	8	0.5
pH (10 wt %)	5-7	5-7
Sulfated ash (% of d.m.)	< 0.2	< 0.2
Heavy metals (ppm of d.m.)	< 0.2	< 0.2
Appearance	white powder	white powder
Taste	neutral	neutral
Sweetness (vs sucrose = 100 %)	10 %	none
Solubility in water at 25 °C (g/L)	120	10
Viscosity in water (5 %) at 10 °C (mPa·s)	1.6	2.4
Heat stability	good	good
Acid stability	fair	good
Functionality in foods	fat replacement	

	body and mouthfeel
	texture improvement
	foam stabilization
	emulsion stabilization
	synergy with gelling agents

*G = glucosyl unit. F = fructosyl unit. d.m. = dry matter.

Physical Properties. Chicory inulin is moderately soluble in water (about 10 % at room temperature) which allows its incorporation in watery systems without precipitation. The viscosity of chicory inulin solutions is rather low, e.g., 1.65 mPa·s at 10 °C for a 5 % dry matter (d.m.) solution and 100 mPa·s for a 30 % d.m. solution [5]. Dissolved inulin exerts a small effect on the freezing and boiling point of water (e.g., 15 % chicory inulin decreases the freezing point by 0.5 °C).

At high concentration (>25 % in water for standard chicory inulin and >15 % for long-chain inulin), inulin has gelling properties and forms a particle gel network after shearing.

The gel strength obtained depends on the inulin concentration and total dry substance content, on the shearing parameters (such as temperature, time, speed, or pressure) and also on the type of shearing device used, but is not influenced by pH (between pH 4 and 9). Electron cryomicroscopy has shown that such an inulin gel is composed of a three-dimensional network of insoluble submicron inulin particles in water. These particles of about 100 nm in size aggregate to form larger clusters having a diameter of 1 to 5 µm. Large amounts of water are immobilized in that network, as determined by NMR experiments. X-ray diffraction confirmed the crystalline nature of the gel particles whereas the starting inulin powder is essentially amorphous.

An inulin gel exhibits a viscoelastic rheological behavior and shows shear-thinning and thixotropic properties. The gel is characterized by a relatively low yield stress (e.g., 1540 Pa for a gel of 30 % standard inulin in water at 25 °C). Inulin also displays synergistic effects with most gelling agents (e.g., gelatin, alginate, κ- and ι-carrageenan, gellan gum, maltodextrin) [10].

Furthermore, inulin improves the stability of foams and emulsions, e.g., in aerated dairy desserts, ice creams, table spreads, and sauces. It can therefore replace other stabilizers in different food products.

Chemical Properties. Under strongly acidic conditions, the β(2 ⟶1) bonds between the fructose units can be (partially) hydrolyzed and fructose is formed. Inulin is stable in applications with a pH > 4. Even at lower pH values, the hydrolysis of inulin is limited to less than 10 % if the products either have a high dry substance content (>70 %), are stored at a low temperature (<10 °C), or have a short shelf-life.

[Top of Page]

4. Physiological Properties

Inulin behaves like a bulk ingredient, contributing to body and mouthfeel. It is often used in combination with intense sweeteners to provide a rounder mouthfeel and a better-sustained flavor with reduced aftertaste.

Inulin provides several interesting nutritional benefits to animals and humans (see Table 2) [11]. Thanks to its β (2 ⟶1) bonds, which the digestive enzymes of humans cannot hydrolyze, inulin passes through the mouth, the stomach, and the small intestine without undergoing any significant change and without being absorbed [12]. Inulin thus enters the colon almost quantitatively and there it is completely metabolized by the intestinal bacteria [13], mainly to short-chain fatty acids (SCFA), bacterial biomass and gases. Only the SCFA contribute to the host's energy metabolism, which explains the reduced caloric value of inulin (between 1 and 1.5 kcal/g). Inulin has no influence on blood glucose and insulin levels when ingested orally and it has been known as a food for diabetics since the beginning of the 20th century [4], [14].

Table 2. Physiological properties of inulin

Strong evidence:
Nondigestibility and low caloric value (1–1.5 kcal/g)
Suitable for diabetics
Soluble dietary fiber
Stool bulking effect: increase in stool weight and stool frequency, relief of constipation
Modulation of the gut flora composition, stimulating beneficial bacteria (*Bifidobacteria*) and repressing harmful ones (*Clostridia*): prebiotic / bifidogenic effect
Improvement of calcium (and magnesium) bioavailability
Promising evidence:
Reduction of serum and liver triglycerides and serum insulin levels
Reduction of colon cancer risk (in animal models)

Modulation of immune response and resistance

Protection against intestinal disorders and infections

Improved well-being

Inulin is a soluble dietary fiber [15] which induces typical effects on the gut function, such as a reduction of the intestinal pH, a relief of constipation, and an increase in stool weight and frequency (or bulking effect). Its fecal bulking effect is similar to that of other soluble fibers such as pectin and guar gum [16]: each gram ingested increases the fecal wet weight by about 2 g. Inulin also has modulating effects on the lipid metabolism, e.g., by reducing serum and liver triglycerides [19], [20].

In the colon, inulin selectively promotes the growth and metabolic activity of beneficial bacteria, mainly *Bifidobacteria*, while repressing harmful ones, e.g., *Clostridia* (prebiotic or bifidogenic effect, ⟶ Health Value Added Foods – Prebiotics) [17], [21], [22]. A daily intake of 5 to 10 g of inulin is sufficient to significantly enhance *Bifidobacteria* in humans [23].

Recent research indicates that inulin also has a significant chemopreventive potential [24]. Indeed, it can prevent the formation of precancerous lesions and tumors in the colon of rats. Long-chain inulin is more effective than shorter-chain fructans. Synergistic effects were observed when inulin and *Bifidobacteria* were administered together as a synbiotic [25]. It was furthermore demonstrated that inulin inhibits the development of cancer cells transplanted in the thigh or in the peritoneum of mice [26].

Inulin also increases the intestinal absorption of calcium, iron, and magnesium as well as the bone mineral density in rats [27], [28]. Studies in human volunteers also show an increased calcium absorption upon administration of inulin [29], [30], [31].

[Top of Page]

5. Resources and Raw Materials

After starch, fructans are the most abundant nonstructural polysaccharides found in nature. They are present in a wide variety of plants and in some bacteria. Most literature on the natural occurrence of fructans does not differentiate, however, between levan and inulin.

Fructan-producing plants are commonly present among the grasses (1200 species), whereas 15 % of the flowering plants produce fructans in significant amounts. Fructan-producing plants are widely spread within the *Liliaceae* (3500 species) and most frequently among the *Compositae* (25 000 species) [32]. A plant with a special status is the *Agave Azul Tequila Weber (Liliaceae)*: Tequila is an alcoholic drink made by fermentation from highly branched fructan present in this plant.

Inulin-containing plants that are commonly used in the human diet belong mainly to either the *Liliaceae* (leek, onion, garlic, and asparagus) or the *Compositae* (Jerusalem artichoke, chicory, and yacon) as shown in Table 3 [33].

Table 3. Inulin content (% of fresh weight) of plants that are commonly used in human nutrition

Source	Edible parts	Dry solids content	Inulin content
Artichoke	leave-heart	14–16	3–10
Banana	fruit	24–26	0.3–0.7
Barley	cereal	NA	0.5–1.5*
Burdock	root	21–25	3.5–4.0
Camas	bulb	31–50	12–22
Chicory	root	20–25	15–20
Dandelion	leaves	50–55*	12–15
Garlic	bulb	40–45*	9–16
Jerusalem artichoke	tuber	19–25	14–19
Leek	bulb	15–20*	3–10
Murnong	root	25–28	8–13
Onion	bulb	6–12	2–6
Rye	cereal	88–90	0.5–1*
Salsify	root	20–22	4–11
Yacon	root	13–31	3–19

NA: figures not available.
*Estimated.

Bacterial inulin is produced by *Streptococcus mutans*, the spores of *Aspergillus sydowi* and *Lactobacillus reuteri*. A high

degree of branching (≥15 %) is observed in these fructans.

Given their high inulin content (>10 %), dahlia, Jerusalem artichoke, and chicory (*Cichorium intybus*) have been considered as candidates for industrial production of inulin [34].

5.1. Dahlia

Many dahlia cultivars are available, but they have all been selected for their flowers, rather than for inulin production. The tuberous roots can be propagated only if attached to a piece of stem tissue. When propagated from seed, sowing has to be delayed until late spring, given dahlia's extreme sensitivity to frost. Mechanical harvesting of the tubers is feasible only on sandy grounds. Although the DP of dahlia inulin is higher than that of chicory inulin, inulin yield from dahlia is only half that of chicory (see Table 4). For all these reasons, dahlia does not appear as an interesting crop for inulin production.

Table 4. Yields and composition of Dahlia, Jerusalem artichoke and Chicory

		Dahlia	Jerusalem artichoke	Chicory
Roots or tubers (t/ha)				
	average	25	42	43.5
	variation		35–60	25–75
d.m. (%)				
	average	18	22	22.5
	variation	15–22	19–25	20–25
Inulin (%)				
	average	11	16	16.5
	variation	10–12	14–18	15–18
Inulin (t/ha)		2.5–3	4.5–8.5	5–11
Mean DP		13–20	6–10	10–14

d.m. = dry matter. DP = degree of polymerization.

5.2. Jerusalem Artichoke

Jerusalem artichoke shows a rather high inulin content (14–19 %). It is not recommended to cultivate Jerusalem artichoke in clay-soil as the tubers are small and irregular, hence bringing a lot of soil attached to them. Although Jerusalem artichoke is frost tolerant, its inulin metabolism is very sensitive to cold. As Jerusalem artichoke inulin has only 20 % of chains with a degree of polymerization longer than 10, it is unsuited for several food and non-food applications (see Table 4).

5.3. Chicory

Chicory is a biennial plant. During the first season, the plants remain in the vegetative phase and make only leaves, tap roots, and fibrous roots. The roots look like small oblong sugar beets. The inulin content is high (16–18 %) and fairly constant from year to year for a given region. Yields in tons of roots per hectare (around 45 t/ha) show a wider variation (Table 4). In order to keep the chicory culture healthy a strict crop rotation is imposed (once every five years). Industrially, inulin is produced from chicory. The chicory used for inulin production (*Cichorium intybus*) is of the same type as the one used for the production of a coffee substitute (after roasting).

[Top of Page]

6. Production

The production of chicory inulin consists of two steps. The first step includes the extraction and primary purification, resulting in a raw impure syrup. The second step is the refining phase, which results in a purified commercial end product [5], [35].

6.1. First Step

The first process phase is virtually similar to sugar beet processing. Roots are harvested and stored in piles on the field. To minimize losses, no more than 7 days should elapse between harvest and delivery to the factory. The roots are then transported to the factory by truck, weighed, and carefully stored on the court. From there, they are transported with a stream of water inside the factory, where they are washed and sliced. Raw inulin is extracted from the resulting “chips” with hot water in a countercurrent diffuser. The leached “chips” are dried and sold as animal feed. A first purification step is applied to the extraction juice by liming and carbonation at high pH. The resulting $CaCO_3$ sludge precipitates easily.

Peptides, some anions, degraded proteins, and colloids are trapped in its flocs. This generates a foam-type product, which is used by the farmers to improve the soil structure, as it is rich in calcium and organic matter.

6.2. Second Step

The raw juice is further refined using cationic and anionic ion exchange resins for demineralization and active carbon for decolorization. This technology is comparable to the one used in starch processing. After demineralization and

decolorization, the juice is passed over a 0.2 µm filter for sterilization, before being evaporated and spray-dried. The resulting standard inulin (RAFTILINE ST, ORAFTI) has an average chain length which reflects the original DP distribution present in the chicory root (between 10 and 12). A special long-chain grade inulin (RAFTILINE HP, ORAFTI), with an average DP >23 is also available. It is made by physical elimination of the small DP fraction [36].

6.3. Environmental Aspects

The chemicals used for regeneration of the ion exchangers are not NaOH and HCl as usual, but NH_3 and H_2SO_4. The advantage is that the effluents can be converted into reusable byproducts. At high concentrations, easily crystallizing salts such as $(NH_4)_2SO_4$ and K_2SO_4 precipitate, are separated from the mother liquor and sold commercially as fertilizers. The mother liquor is further evaporated into a stable product that is sold as animal feed because of its high organic matter content. This way the circle between the processing industry and the agriculture is fully closed.

[Top of Page]

7. Uses

7.1. Food Applications

In food and drink products, inulin can be used for either its nutritional advantages or its technological properties, but it is often applied to offer a double benefit: an improved organoleptic quality and a better-balanced nutritional composition. Table 5 gives an overview of the applications of inulin in foods and drinks.

The use of inulin as a dietary fiber ingredient is easy and often leads to an improved taste and texture [37]. Its solubility allows fiber incorporation in watery systems such as drinks, dairy products, and table spreads. Inulin is more and more used in “functional foods”, especially in a whole range of dairy products, as a prebiotic ingredient (⟶ Health Value Added Foods – Prebiotics).

Table 5. Applications of inulin in foods and drinks

Application	Functionality	Dosage level (wt %)
Dairy products (yogurts, cheeses, desserts, drinks)	fat replacement body and mouthfeel foam stabilization fiber and prebiotic	2–10
Frozen desserts	fat replacement texture improvement melting behavior low caloric value	2–10
Table spreads and butter-like products	fat replacement texture and spreadability emulsion stabilization replacement of gelatine fiber and prebiotic	2–10
Baked goods and breads	fiber and prebiotic moisture retention	2–15
Breakfast cereals and extruded snacks	fiber and prebiotic crispness and expansion low caloric value	2–20
Fillings	fat replacement texture improvement	2–30
Salad-dressings and sauces	fat replacement mouthfeel and body emulsion stabilization	2–10
Meat products	fat replacement texture and stability fiber	2–10
Dietetic products and meal replacers	fat replacement synergy with intense sweeteners body and mouthfeel fiber and low calorie	2–15
Chocolate	sugar replacement and	5–30

	fiber and heat resistance	
Tablets	sugar replacement fiber and prebiotic	5–75

Thanks to its specific gelling characteristics, inulin allows the development of low-fat foods without compromising on taste and texture. When the fructan is thoroughly mixed with water or another aqueous liquid, using a shearing device such as a rotor-stator mixer or a homogenizer, a white creamy structure results which can easily be incorporated in foods to replace fat up to 100 % [5], a process patented by RAFFINERIE TIRLEMONTOISE/ORAFTI [9]. In table spreads, both fat and water-continuous, inulin allows the replacement of significant amounts of fat and the stabilization of the emulsion, while providing a short spreadable texture. In low-fat dairy products, such as fresh cheese, cream cheese or processed cheese, the addition of a few percents of inulin gives a creamier mouthfeel and imparts a better-balanced round flavor. Inulin also is destined to be used as a fat replacer in frozen desserts, providing an easy processing, a fatty mouthfeel, excellent melting properties, as well as freeze-thaw stability, without any unwanted off-flavor. Fat-reduced meat products, such as sausages and pâtés with a creamier and juicier mouthfeel and an improved stability due to better water immobilization, can be obtained using inulin as ingredient. The synergistic effect of inulin with other gelling agents constitutes an additional advantage in all these applications.

Special “instant” qualities, which do not require shearing to give stable homogeneous gels also have been developed (using specific spray-drying technology) and patented.

The incorporation of inulin (1 to 3 %) in diet fruit yogurts, possibly through the fruit preparation, improves the mouthfeel, and offers a synergistic taste effect in combination with aspartame and/or acesulfame K [8]. Inulin furthermore increases the stability of foams and mousses: its incorporation at 1 to 5 % into dairy-based aerated desserts improves the processibility and upgrades the quality.

Inulin also has found an interesting application as a low calorie bulk ingredient in chocolate without added sugar, often in combination with a polyol. It is also used as a dietary fiber or sugar replacer in tablets.

7.2. Inulin Derivatives for Non-Food Applications

In the context of product development based on renewable resources, different types of chemical modifications have been performed using inulin as raw material [38]. From the industrial point of view, three reaction types are most interesting: oxidation, carboxymethylation, and hydrophobization.

7.2.1. *Oxidation*

Although inulin can be oxidized in several ways, a one-step reaction using sodium hypochlorite as the oxidant together with sodium bromide as a catalyst is generally used (see Fig. 2) [38-41].

Figure 2. Oxidation of inulin

Under well-controlled conditions (slow addition of the oxidant in 4 to 6 h, constant pH of 9.5, addition of a concentrated NaOH solution) oxidation occurs almost exclusively at the C3-C4 fructose bond, resulting in a ring-opened inulin known as dicarboxyinulin (DCI). Due to this fructose ring-opening and the direct formation of the calcium-complexing oxydiacetate (ODA) structure, DCI shows excellent calcium sequestering properties (SC) (2.0–2.5 mmol Ca/g active substance) [39]. Moreover, due to the linear relationship between the degree of oxidation, the formation of the ODA structure and hence the SC value, it is easy to compromise between the oxidation degree and the biodegradability.

7.2.2. *Carboxymethylation*

Heating an aqueous solution of inulin in the presence of an excess of sodium hydroxide with monochloroacetic acid results in the production of O-(carboxymethyl) inulin (CMI) (see Fig. 3) [42, 43]. The reaction efficiency of this etherification very much depends on parameters such as reaction temperature and water content: lower water content and temperature result in a higher efficiency.

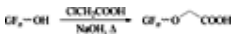

Figure 3. Carboxymethylation of inulin

CMI is a good inhibitor of the crystallization of calcium carbonate. Even at very low concentrations (5–200 ppm), CMI can extend the induction period for crystallization by absorption of the carboxylate groups onto the crystal surfaces.

7.2.3. *Hydrophobization*

Hydrophobization of inulin, e.g., grafting of alkyl chains on the inulin backbone, can be done in two different ways: by esterification or by carbamoylation.

The esterification of inulin is carried out by reaction with acid chlorides or anhydrides; the introduced alkyl chain can range from 1 to 22 carbon atoms (see Fig. 4) [44]. The reaction is usually performed either in solvents such as pyridine, *N,N*-dimethylformamide or dimethyl sulfoxide, or solvent-free, using a kneader or (preferably) an extruder. As catalyst 4-(dimethylamino)pyridine, sodium acetate, or potassium carbonate can be used.

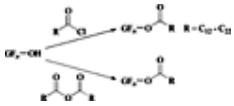

Figure 4. Esterification of inulin

The carbamoylation of inulin (modification of inulin with an alkyl-isocyanate) is done in a polar organic solvent such as *N,N*-dimethylformamide or *N*-methylpyrolidinone, as shown in Figure 5 [45].

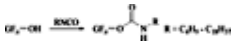

Figure 5. Carbamoylation of inulin

Both the esterified and carbamoylated inulins show interesting interfacial properties. Depending on the length of the alkyl chain and the degree of substitution these inulin derivatives are able to decrease the surface tension of aqueous solutions and the interfacial tension at an oil–water interface [46]. Moreover, the critical micelle concentration (cmc) of hydrophobized inulins is approximately 0.01 % (w/v). The above described tensioactive properties make these molecules suitable for use in detergent applications, cosmetics, or crop protection formulations.

[Top of Page]

8. Analysis

Although several analytical techniques are available for the determination of inulin, no single method gives a complete and quantitative analysis of all the components present. A combination of different methods is therefore often necessary.

8.1. HPLC

Two columns in series in the K^+ form are used (Aminex HPX-87 K^+) for optimal separation. The separation into fructose, glucose, DFA (difructose-dianhydride), GF, F_2, and F_3 is optimal, but further separation into DP3, DP4, and DP ≥ 5 is not very precise. DP3 and DP4 fractions are not pure, but include GF_2 plus F_4 and GF_3 plus F_5, respectively. The fraction of DP ≥ 5 is the integrated sum of DP5 and higher DP molecules. As this fraction might include small oligosaccharides as well as high DP inulins, a fixed response factor can not be determined, which makes this analysis unsuited for quantitative inulin determination. The method is well adapted to evaluate the relative amounts of the different compounds present; especially the amounts of the non-inulin compounds glucose, fructose, and sucrose.

8.2. Gas Chromatography

A high temperature capillary gas chromatographic method has been developed for the quantitative determination of inulin-type oligosaccharides with DP <10 [47]. Sample preparation involves oximetrimethylsilylation (oxymation) and silylation of the extracted sugars. The method is accurate and specific.

Moreover $\beta(2 \longrightarrow 6)$ oligosaccharides can be clearly distinguished from $\beta(2 \longrightarrow 1)$ compounds and GF_n molecules from F_n molecules.

8.3. HPAEC Analysis (Dionex)

A differentiation between GF_n and F_n compounds can be made using High Pressure Anion Exchange Chromatography (HPAEC) which furthermore provides a fingerprint of the molecular mass distribution of inulin. This analytical technique uses a Dionex series 4000 ion chromatograph (Carbo-Pac PA-1 column) coupled with a Pulsed Amperometric Detector (PAD). During the analysis, the carbohydrates are eluted with a NaOH/NaAc gradient [48]. The major drawback of HPAEC-PAD is that it is very difficult to quantify the high DP compounds, due to the lack of appropriate standards and the reduced sensitivity of the PAD detector for high DP polymers.

When coupling HPAEC with a Pulsed Electrochemical Detector (PED), oligomers up to DP17 can be quantified and the weight fraction of each compound present can be calculated [49].

The average degree of polymerization is determined by the fructose to glucose ratio after complete hydrolysis with inulinase (Fructozyme, NOVO) (at pH 4.5 and 60 °C for 30 min): $\mathrm{D\bar{P}n} = \mathrm{F/G} + 1$.

8.4. Permethylation

The type of linkage and the occurrence of branching are checked by permethylation followed by reductive cleavage and in situ acetylation [50]. The reaction products are analyzed using capillary gas chromatography (CGC).

8.5. Quantitative Determination

The official AOAC method no. 997.08 was developed to quantitatively measure inulin, e.g., in food [51]. An overview of the method is given in Figure 6.

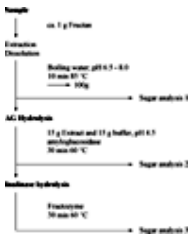

Figure 6. Flow diagram of the enzymatic fructan determination method (AOAC no. 997.08)

If inulin is the only compound present in the sample, the method consists only of steps 1 and 3. The inulin is extracted from a substrate at 85 °C for 10 min; part of the extract is put apart for determination of free fructose, glucose, and sucrose by any reliable chromatographic method available (HPLC, HGC, or HPAEC-PAD), the other part is submitted to an enzymatic hydrolysis, identical to the one described above for DP determination. After this hydrolysis step, resulting fructose, and glucose are determined again by chromatography. By subtracting the initial glucose, fructose, and sucrose contents from the final ones, the following formula can be applied:

$$\text{Inulin} = k(\text{G}_{\text{inu}} + \text{F}_{\text{inu}})$$

where k (<1) depends on the DP of the inulin analyzed and corrects for the water gain after hydrolysis. G_{inu} and F_{inu} are glucose and fructose, respectively, strictly originating from inulin.

If a complex sample needs to be analyzed, as is often the case when dealing with food products, an amyloglucosidase treatment must be included before the last step and an extra sugar analysis performed to avoid overestimation of the glucose originating from starch and maltodextrins present.

Although AOAC method 997.08 is very reliable, it is very labor intensive and requires the use of chromatographic apparatus. AOAC method no. 999.03, based on the use of enzymes for hydrolysis and sugars determinations, just requires a spectrophotometer and some other standard lab equipment [52]. This method is reliable for the determination of inulin, but it cannot be used for the determination of oligosaccharides obtained by hydrolysis of inulin, as F_n-type compounds are significantly underestimated.

[Top of Page]

9. Legal Aspects

Inulin is legally classified as food or food ingredient, and not as additive, in all countries where it is used. As a consequence, inulin is not listed as “accepted food additive” in the standard positive lists, e.g., from the European Union or from Codex Alimentarius. In the USA, inulin was declared “Generally Recognized As Safe” in 1992.

Inulin is classified as ‘dietary fiber’ in all European countries and in most other countries. It also complies with the Codex Alimentarius definition of Dietary Fiber (Codex Guidelines on Nutrition Labelling CAC/GL 2-1985, Rev. 1 - 1993).

[Top of Page]

10. Economic Aspects and Trade Names

Currently, inulin is produced industrially by three companies, all of them using chicory as raw material: ORAFTI, Belgium, a subsidiary of RAFFINERIE TIRLEMONTOISE (trade names RAFTILINE for food applications, RAFTIFEED for animal feed and INUTEC for non-food uses), WARCOING, Belgium (trade name FIBRULINE) and SENSUS, The Netherlands, a subsidiary of COSUN (trade name FRUTAFIT).

Chicory processing has evolved from a laboratory and pilot plant project to a full industrial scale. While the production of inulin as such is not regulated by quota, the production of chicory fructose (obtained by hydrolysis of inulin, see ⟶ Fructose) is. As these quota are officially attributed by the European Union, they give an idea of the production scale [5].

The fact that the land surface assigned to chicory for inulin/fructose production has evolved from a few hundred hectares in 1990 to more than 20 000 ha in 2000 is another indication for the growth of the inulin-related business.

[Top of Page]

References

1. V. Rose, Über eine eigenthümliche vegetablische Substanz, *Gehlens Neues Algem. Jahrb. Chem.* **3** (1804) 217.
2. T. Thomson: *A System of Chemistry*, 5th London edition, vol. **4**, Abraham Small, Philadelphia 1818, p. 65.
3. J. Sachs, Über die Sphärokristalle des Inulins und dessen mikroskopische Nachweisung in den Zellen, *Bot. Ztg.* **22** (1864) 77.
4. H. Root, M. Baker, Inulin and artichokes in the treatment of diabetes, *Arch. Intern. Med.* **36** (1925) 126–145. Links
5. A. Franck, L. De Leenheer: “Inulin”, in A. Steinbuchel (ed.): *Biopolymers*, vol. **6**, VCH Weinheim, Germany 2002, pp. 439–479.
6. A. L. Watherhouse, N. J. Chatterton: “Glossary of fructan terms”, in M. Suzuki, N. J. Chatteron (eds.): *Science and*

Technology of fructans, CRC Press, Florida 1993, pp. 2–7.

7. L. De Leenheer, H. Hoebregs, *Progress in the elucidation of the composition of Chicory inulin*, Starch **46** (1994) 193–196.
8. M. Wiedmann, M. Jager, *Synergistic sweeteners, Food Ingredients and Analysis International*, Nov.–Dec. (1997) 51–56.
9. Tiense Suikerraffinaderij, Compositions having a creamy structure and containing fructan, preparation method therefor and uses thereof, Patent Application WO 93 06744, 1993 (A. Frippiat, G. Smits).
10. Tiense Suikerraffinaderij, Inulin based hydrocolloid compositions, Patent Application EP 867 470, 1998 (A. Frippiat).
11. J. Van Loo, J. Cummings, N. Delzenne, H. Englyst, A. Franck, M. Hopkins, N. Kok, G. Macfarlane, D. Newton, M. Quigley, M. Roberfroid, T. van Vliet, E. van den Heuvel, Functional food properties of non-digestible oligosaccharides: a consensus report from the ENDO project (DGXII AIRII-CT94-1095), *Br. J. Nutr.* **81** (1999) 121–132. Links
12. L. Ellegard, H. Andersson, I. Bosaeus, Inulin and oligofructose do not influence the absorption of cholesterol, and the excretion of cholesterol, Fe, Ca, Mg and bile acids but increase energy excretion in man. A blinded, controlled cross-over study in ileostomy subjects, *Eur. J. Clin. Nutr.* **51** (1997) 1–5. Links
13. M. Roberfroid, G. R. Gibson, N. Delzenne, Biochemistry of oligofructose, a non-digestible fructo-oligosaccharide: an approach to estimate its caloric value, *Nutrition Reviews* **51** (1993) 137–146. Links
14. A. Beringer, R. Wenger, Inulin in der Ernährung des Diabetikers, *Deutsch. Zeitschr. F. Verdauungs-u. Stoffwechselkrankh.* **15** (1955) 268–272.
15. L. Prosky, Inulin and Oligofructose are Part of the Dietary Fiber Complex, *J. AOAC Int.* **82** (1999) 223–226. Links
16. M. B. Roberfroid: “Health benefits of non-digestible oligosaccharides,” in Kritchevsky and Bonefield (Eds.): *Dietary Fiber in Health and Disease*, Plenum Press, New York 1997, pp. 211–219.
17. B. Kleessen, B. Sykura, H. J. Zunft, Effect of inulin and lactose on fecal microflora, microbial activity and bowel habit in elderly constipated persons, *Am. J. Clin. Nutr.* **65** (1997) 1397–1402. Links
18. E. Den Hond, B. Geypens, Y. Ghoos, Effect of high performance chicory inulin on constipation, *Nutrition Research* **20** (2000) 731–736. Links
19. N. M. Delzenne, C. Daubioul, A. Neyrinck, M. Lasa, H. S. Taper, Inulin and Oligofructose modulate lipid metabolism in animals: review of biochemical events and future prospects, *Br. J. Nutr.* **87** (2002) Suppl.2 S225–S259.
20. K. G. Jackson, G. R. J. Taylor, A. M. Clohessy, C. M. Williams, The effect of the daily intake of inulin on fasting lipid, insulin and glucose concentrations in middle-aged men and women, *Br. J. Nutr.* **82** (1999) 23–30. Links
21. G. R. Gibson, X. Wang, *Bifidogenic properties of different types of fructo-oligosaccharides, Food Microbiology* **11** (1994) 491–498.
22. G. R. Gibson, E. R. Beatty, X. Wang, J. H. Cummings, *Selective stimulation of bifidobacteria in the human colon by oligofructose and inulin, Gastroenterology* **108** (1995) 975–982.
23. M. Roberfroid, J. Van Loo, G. Gibson, *The Bifidogenic nature of chicory inulin and its hydrolysis products, J. Nutr.* **128** (1998) 11–19.
24. D. S. Reddy, R. Hamid, C. V. Rao, *Effect of dietary oligofructose and inulin on colonic preneoplastic aberrant crypt foci inhibition, Carcinogenesis* **18** (1997) 1371–1374.
25. I. R. Rowland, C. J. Rummey, J. T. Coutts, L. Lievense, *Effects of Bifidobacterium longum and inulin on gut bacterial metabolism and carcinogen-induced aberrant crypt foci in rats, Carcinogenesis* **19** (1998) 281–285.
26. H. Taper, N. Delzenne, M. B. Roberfroid, Growth inhibition of transplantable mouse tumours by non-digestible carbohydrates, *Int. J. Cancer* **71** (1997) 1109–1112. Links
27. N. Delzenne, J. Aertssens, N. Verplaetse, M. Roccaro, M. Roberfroid, Effect of fermentable fructo-oligosaccharides on energy and nutrients absorption in the rat, *Life Science* **57** (1995) 1579–1587.
28. M. B. Roberfroid, J. Cumps, J. P. Devogelaer, Dietary Chicory Inulin increases Whole-Body Mineral Density in Growing Male Rats, *J. Nutr.* **132** (2002) 3599–3602. Links
29. C. Coudray, J. Bellanger, C. Castiglia-Delavaud, C. Rémésy, M. Vermorel, Y. Rayssignuier, Effects of soluble or partly soluble dietary fibres supplementation on absorption and balance of calcium, magnesium, iron and zinc in healthy young men, *Eur. J. Clin. Nutr.* **51** (1997) 375–380. Links
30. I. J. Griffin, P. M. Davila, S. A. Abrams, Non-digestible oligosaccharides and calcium absorption in girls with adequate calcium intakes, *Br. J. Nutr. 87 Suppl.* **2** (2002) S187–S191.
31. A. Franck, Prebiotics stimulate calcium absorption: a review, *Milchwissenschaft* **53** (1998) 427–429. Links
32. G. A. Hendry, R. K. Wallace: “The origin, distribution and evolutionary significancy of fructans”, in M. Suzuki, N. J. Chatteron (eds.): *Science and Technology of fructans*, CRC Press, Florida 1993, pp. 119–139.
33. J. Van Loo, P. Coussement, L. De Leenheer, H. Hoebregs, G. Smits, On the presence of inulin and oligofructose as natural ingredients in the Western diet, *Crit. Rev. in Food Sci. Nutr.* **35** (1995) 525–552.
34. W. J. M. Meyer, E. W. J. M. Matthysen, G. E. L. Borm: “Crop characteristics and inulin production of Jerusalem Artichoke and Chicory”, in A. Fuchs (ed.): *Inulin and inulin-containing crops, Studies in Plant Science*, 3rd ed., Elsevier Science Publishers, 1993, pp. 29–44.
35. L. De Leenheer: “Production and use of inulin: Industrial reality with a promising future”, in H. Van Bekkum, H. Röper, F. Voragen (eds.): *Carbohydrates as organic raw materials III*, VCH Publ. Inc., New York 1996, pp. 67–92.
36. Tiense Suikerraffinaderij, Fractionated polydisperse compositions, Patent Application WO 96 01849, 1994 (G. Smits, L. Daenekindt, K. Booten).
37. A. Franck, Technological functionality of inulin and oligofructose, *Br. J. Nutr.* **87 Suppl. 2** (2002) S287–S291. Links
38. C. Stevens, A. Meriggi, K. Booten, Chemical modification of inulin, a valuable renewable resource and its industrial applications, *Biomacromolecules* **2** (2001) 1–16. Links
39. A. Besemer, H. van Bekkum, The relation between calcium sequestering capacity and oxidation degree of dicarboxy-

starch and dicarboxy-inulin, *Starch/Stärke* **46** (1994) 419–422.

40. TNO, Method for preparation of calcium-binding polycarboxy compounds based on polysaccharides and replacements for phosphates in detergents, based on these polycarboxy compounds, PCT WO 91/17189, 1991 (A. Besemer, H. Van Bekkum).
41. D. Verraest, J. Peters, H. van Bekkum, Oxidation and carboxymethylation of sucrose and inulin, *Zuckerindustrie* **120** (1995) 799–803. Links
42. D. Verraest, J. Peters, J. Batelaan, H. van Bekkum, Carboxymethylation of inulin, *Carbohydrate Research* **271** (1995) 101–112. Links
43. Akzo Nobel, Carboxymethyl inulin, PCT WO 95/15984, 1995 (D. Verraest, J. Batelaan, J. Peeters, H. van Bekkum).
44. Südzucker, Aliphatic Carboxylate esters of inulin, US Patent US 5 877 144, 1999 (S. Ehrhardt, A. Haji Begli, M. Kunz, L. Sheiwe).
45. Tiense Suikerraffinaderij, Surface-active alkyl-urethanes of fructans, PCT WO 99/64549, 1999 (C. Stevens, K. Booten, L. Laquiere, L. Daenekindt).
46. C. Stevens, A. Meriggi, M. Peristeropoulou, P. Christov, K. Booten, B. Levecke, A. Vandamme, N. Pittevils, T. Tadros, Polymeric surfactants based on inulin, a polysaccharide extracted from chicory. 1. Synthesis and interfacial properties, *Biomacromolecules* **2** (2001) 1256–1259. Links
47. D. Joye, H. Hoebregs, Determination of oligo-fructose, a soluble dietary fiber, by high-temperature capillary gas chromatography, *J. AOAC Int.* **83** (2000) 1020–1025. Links
48. N. J. Chatteron, P. A. Harrison, W. R. Thornley, J. H. Bennet: “Separation and quantification of fructan (inulin) oligomers by anion exchange chromatography”, in A. Fuchs (ed.): Inulin and Inulin-containing crops, 3rd ed., Elsevier Science Publishers, 1993, pp. 93–99.
49. J. W. Timmermans, M. B. van Leeuwen, H. Tournois, D. de Wit, J. F. G. Vliegenthart: *Proceedings of the fourth seminar on inulin*, Wageningen, 1993, pp. 12–15.
50. I. Ciucanu, F. Kerek, A simple and rapid method for the permethylation of carbohydrates, *Carbohydr. Res.* **131** (1984) 209–217. Links
51. H. Hoebregs, Fructans in foods and food products, ion-exchange chromatographic method: Collaborative study, *J. AOAC Int.* **80** (1997) 1029–1037. Links
52. B. McCleary, A. Murphy, D. Mugford, Measurement of total fructan in foods by enzymatic/ spectrophotometric method: Collaborative study, *J. AOAC Int.* **83** (2000) 356–364. Links
53. P. A. A. Coussement, Inulin and Oligofructose: Safe intakes and Legal Status, *J. Nutr.* **129** (1999) 1412S–1417S. Links

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Freeze-Drying

Standard Article

Georg-Wilhelm Oetjen[1]
[1]Lübeck, Germany

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.h12_h01
Article Online Posting Date: January 15, 2004

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. Goals of Industrial Freeze-Drying
1.2. Steps of the freeze-drying process
1.3. Chemical Engineering Tasks in Industrial Freeze-Drying
2. Process Steps
2.1. Freezing
2.2. Main Drying
2.3. Transition from Main to Secondary Drying (SD)
2.4. Secondary Drying
2.5. Packing and Storage
3. Temperatures and Pressures
3.1. Barometric Temperature Measurement
3.2. Pressure During Secondary Drying
3.3. Water-Vapor Flow
3.4. Residual Moisture Content
3.5. Uniformity of Temperatures, Pressures, and Product
4. Structures of Frozen and Dried Products
4.1. Influence of Freezing Method and Excipients
4.2. Methods of Structure Analysis
5. Industrial Freezing and Freeze-drying Installations
5.1. For Pharmaceutical Products
5.2. Regulatory Requirements for Pharmaceutical Plants
5.3. Food Plants
5.4. Process Automation
5.4.1. Prerequisites for Process and Plant Automation
5.4.2. Control of the Process and the Plant by Thermodynamic Data Measured During the Process: Thermodynamic Lyophilization Control
5.4.3. Transfer of Pilot Plant Data to Production
6. Some Examples of Freeze-drying Processes
6.1. For Pharmaceuticals
6.1.1. Process Reproducibility
6.1.2. Shorter Drying Cycle
6.1.3. Selection of Exipients
6.2. For Food
6.3. For Metal Oxides, Ceramic Powders
7. Economic and Social Aspects, Outlook

[Top of Page]

1. Introduction

1.1. Goals of Industrial Freeze-Drying

The freeze drying of pharmaceutical and biological products is performed for one or more of the following reasons:

1. The ingredients of the formulation are not stable in the liquid state, and other methods of water removal destroy or reduce the active ingredient
2. The amount of active ingredients is very small, and the dosing of liquids can be better controlled than that of a powder
3. Sterilization of the vials before filling and filling into the final container minimizes handling and reduces possible contamination
4. The desired structure of the product can only be achieved by solidifying it and removing the solvent in this phase

The freeze drying of food aims at convenient, natural, storage-stable products. The weight reduction to mostly 10 or 20 % is of interest only in some cases. The main freeze-dried food products are coffee extract, spices, some fruits, and meat. Marketing and economy are the deciding criteria for freeze-dried food.

Freeze-drying of collagen is carried out on an industrial scale; oxides and ceramics are more or less in a development phase.

1.2. Steps of the freeze-drying process

The first step is to transform the solvent, which in all cases in this article is water, into a form with drastically reduced mobility of its molecules. This can be in the form of ice crystals, an amorphous phase, or as residual water in a highly viscous concentrate produced by freeze-concentration. Ice crystals are often undesired, e.g., in organic cells and their suspensions, in most protein solutions, and in several food products. The solidification of the product into a structure, its “freezing”, is a decisive step in freeze-drying. The structure of the frozen product decides:

- Whether a product can be dried at all
- The maximum temperature at which the frozen water can be sublimed
- The pressure at which the remaining water must be desorbed
- The minimum water content that can be achieved at the end of drying
- At which temperature and how long the product can be safely stored

These five points of course also depend on the ingredients of the product. Ingredients and freezing method decide the drying process and the storage capability.

The second step is the sublimation of the ice, known as main drying (MD). Theoretically, this is possible at atmospheric pressure, but three facts make such a process technically impossible, in spite of many trials and ideas:

1. 1 kg of ice sublimed at 0.6 mbar has a volume of ca. 2000 m^3. The air volume to transport this water vapor must be in the ratio of 0.6 to 1000 mbar, i.e., ca. 3.3×10^6 m^3 per kilogram of ice.
2. If only the diffusion of water vapor in quiescent air is used and the distance between ice and condenser is 100 cm, only 4×10^{-2} g m^{-2} h^{-1} can be transported.
3. If the product to be dried is mixed with a strong water vapor absorber, the distance between ice and absorber is very small, but firstly an acceptable absorber must be found, and secondly product and absorber must be separated quantitatively after drying.

Therefore, only vacuum processes are discussed in this article.

The third step is the desorption of water from the solid, called secondary drying (SD). Normally this process starts at a water content (called RM) of ca. 10 % (or 0.1 g water/g solids) and ends between 3 and 0.3 %, depending on the product.

Step four is the packing of the product in a packaging tight to water vapor and often to O_2. Vials with product are placed with a stopper in a semi-open position on the shelves of the freeze-dryer, and at the end of drying the stoppers are pushed into the closed position, either under vacuum or after venting the chamber with a defined gas of a prescribed pressure. Product in open trays, as most food, are vented in the chamber with a dry gas and packed in containers with as little uptake of water vapor or oxygen as possible from the air.

Step five is the storage of the product, which lasts 100 or normally 1000 times longer than the drying process. Long-term structural changes and/or chemical reactions will influence this part of the process. Generally the storage temperature and the water content at the beginning of storage will dominate over the product behavior.

1.3. Chemical Engineering Tasks in Industrial Freeze-Drying

Freeze-drying is a complex sequence of steps based on several scientific and technological disciplines: From biology to statistics, from pharmacy to thermodynamics, from mechanical engineering to production control and automation. In reviewing several thousand freeze-drying publications over the years, it becomes evident that the chemical engineering aspects are very often not recognized by pharmacists, biologists, and mechanical or production engineers. Often one can read excellent papers about structures of proteins, survival rates of cells or sugars as excipients, but the information on the freeze-drying process is limited to “frozen and freeze-dried”. It should therefore be a major task for the chemical engineers or for the process or production development group to inform all parties concerned about the unavoidable requirements of an industrial freeze-drying process. This can literally save millions in costs and months in time, as a few examples

illustrate:

1. In an automatically loaded and unloaded plant temperature sensors can not be used; freezing, main drying, and secondary drying processes must be controlled by other methods. If this is not realized during product and process development, a large part of the development might have to be repeated with other temperature measurement methods.
2. The selection of an optimum vial can easily influence the main drying time by a factor of two, some time up to a factor of four; the main drying time could be, e.g., 15 h, 30 h, or even 60 h for the same product, without changing anything but the vial. Changing the type of vial later will require a substantial amount work and time, or the cycles will be unnecessary long.
3. The uniformity of a large charge is a must in production. Influences of wall and door temperatures may not be recognized in laboratory or pilot plants. They are present but are not documented sufficiently to be transferred to a production plant. Substantial studies at the production plant may be required to achieve uniform production in the process time of the pilot plant. Details of such influences are described in Section Uniformity of Temperatures, Pressures, and Product.

Another task for chemical engineering is to study the possibility of partially or totally automating the total freeze-drying process. Sooner or later this will become a requirement for the following reasons: In the pharmaceutical industry freeze-drying plants are becoming larger and larger (50 000 to 100 000 vials per charge), and the value of the product per vial is increasing; one charge can represent often several million euros. Human errors, which inevitably occur, must be avoided as much as possible, and humans are the most frequent danger to sterility in the production area; their numbers should be reduced as much as possible. In freeze-drying of food the cost of operators is an important factor in the total cost; such plants are mostly run continuously seven days a week, and the number of operators should be minimized, as should their errors. The automation of production freeze-drying plants must be considered at all times. Automation adds to the process requirements as shown in example 1 above and influences several other steps, as shown in Section Process Automation.

[Top of Page]

2. Process Steps

2.1. Freezing

The result of cooling a solution or suspension below 0 °C depends on the type and the amount of nuclei available for crystallization, the degree of supercooling, the freezing rate (°C/min), the mobility of the water molecules in the product, and the end temperature of freezing. These influences are partially dependent on each other, but in simplified and general terms, the following cases can be distinguished:

- Slow freezing (rate < 0.5 °C/min): large ice crystals, mostly equilibrium conditions, freeze-concentration of solids most likely, water crystallizes to a maximum extent.
- Quick freezing (rate 0.5–2 °C/min): these rates can be achieved on the shelf of a freeze-drying plant, 2 °C/min only under special conditions.
- Fast freezing (5–300 °C/min): this can only be achieved with liquid nitrogen (LN_2) and above about 100 °C/min with droplet freezing in LN_2. Mostly small ice crystals are formed; supercooling can be substantially extended (e.g., 10–20 °C); freeze-concentration only in small regions distributed over the product; some water molecules have no time to move to a crystal or nucleus, and this water can behave mechanically like a solid, but during rewarming it will be in the liquid phase (unfrozen water).

In pure water without any nuclei for crystallization ice forms at ca. –40 °C. This process is called homogeneous crystallization, while crystallization in the presence of third particles is called heterogeneous crystallization.

In all cooling processes a situation in the product is reached in which the water molecules can no longer move to the nuclei or the ice crystals, since the mobility of the water molecules has become too small in the highly viscous surroundings. The amount of water still not crystallized at a typical temperature for the product is called unfreezable water (UFW). Figure 1 shows the phase diagram of water–glycerol: The three phases of ice which exist under normal pressure as a function of temperature are shown, and the time it takes to change from one type to another. If water vapor is condensed on a cold surface in a very thin film, amorphous ice is formed; it remains stable at –160 °C for a long time; as shown in Figure 1, the change from amorphous to cubic ice will take about 5×10^5 min, or more than a year. The rate of change depends very much on the temperature: at –135 °C the same change takes only 1 min. This change is called devitrification. At –125 °C the change from cubic to hexagonal ice takes about 1000 h, while at –65 °C only hexagonal ice is stable. This behavior of pure water changes if water solutions, suspensions in water, and mixtures with water are studied, as it is the case for virtually all products to be freeze-dried. Luyet [2] shows diagrammatically (Fig. 2) how various phase changes take place at different glycerol concentrations. At 60 % glycerol devitrification takes place at ca. –115 °C, and the devitrification temperature increases with increasing glycerol concentration to ca. –85 °C. However, such high concentrations of glycerol can not be used to freeze dry organic substances. The devitrification temperature T_g depends strongly on the content of unfrozen water; incompletely frozen products have a very low T_g (e.g., –85 °C), while this temperature can be much higher (e.g., –58 °C), if all water (apart from the UFW) is crystallized; this temperature is normally indicated as $T_{g'}$.

Figure 1. Phase diagram water–glycerin

On the left-hand side the dependence of the phase transformation time on the ice temperature is shown: At –140 °C

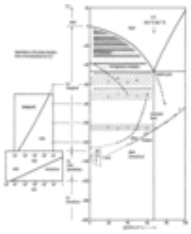

amorphous ice transforms into cubic ice in ca. 10 min. (Fig. 8 from [1]).

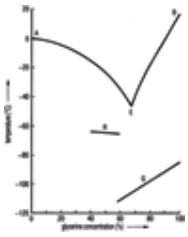

Figure 2. Temperature as a function of the concentration of water–glycerol mixture at which phase transformations occur. (Fig. 14 from [2])

Definitions by Luyet: AE: formation of small crystals or molecular groups, E: eutectic point, EB: formation of clusters, R: eruptive recrystallization, G glass transition.

The freezing time t_f can be estimated by Equation (1) [3]

$$t_f = \Delta J / \Delta T \rho_g \left(d^2 / 2\lambda_g + d / K_{su} \right) \quad (1)$$

where t_f is the freezing time, ΔJ the enthalpy difference between the initial freezing point and the final temperature, ΔT the difference in temperature between the freezing point and the cooling medium, d the thickness of the product parallel to direction of prevailing heat transfer, ρ_g the density of the frozen product, λ_g the thermal conductivity of the frozen product, and K_{su} the surface heat-transfer coefficient between cooling medium and the freezing zone.

The importance of vial selection, mentioned in Section Chemical Engineering Tasks in Industrial Freeze-Drying, also applies to freezing. The only two factors which can be influenced in Equation (1) are ΔT and K_{su}. K_{su} can vary by a factor of two (see Table 1), but lower values (e.g., 100) are also found. Quick freezing rates, one of the most important factors to reduce freeze-concentration and large crystals, can only be achieved by means of high K_{su} and large ΔT. The latter is limited by the brine viscosity/temperature in the shelves, e.g., to a minimum of –75 to –80 °C. This representation assumes that the heat transfer conditions from the brine to the shelf are optimum and more favorable than those from the shelf to the vial or tray (see Section Uniformity of Temperatures, Pressures, and Product).

Table 1. Surface heat-transfer coefficient, total heat-transfer coefficient, and thermal conductivity [29]

K_{su} from gases to a solid surface (kJ m^{-2}h^{-1} °C^{-1})	
free convection	17–21
laminar flow 2 m/s	50
laminar flow 5 m/s	100
K_{su} between the shelf of a freeze-drying plant and a product in vials or trays during freezing (kJ m^{-2} h^{-1} °C^{-1})	200–400
K_{su} between a liquid and a "solid" surface (kJ m^{-2} h^{-1} °C^{-1})	
LN_2 by drops on the product [33]	900
product drops in LN_2	up to 300
K_{tot} between the shelf of a freeze-drying plant and the sublimation front in the product contained in vials or trays under vacuum (kJ m^{-2} h^{-1} °C^{-1})	60–130
λ, thermal conductivity (kJ m^{-2} h^{-1} °C^{-1})	
λ_g, frozen product (ice) [34]	5.9–6.3
λ_{tr}, dry product [35-38]	0.059–0.29

The influence of the freezing rate on a 0.9 % NaCl solution can be seen by an electrical resistance measurement of the sample during freezing and warming in Figure 3 (3 °C/min) and Figure 4 (1 °C/min): at 3 °C/min the expected change at –21.7 °C is visible but mostly delayed to –30 °C (no equilibrium condition), during warming, as best seen in the first derivative, the mobilty of the ions increases relatively slowly over a temperature range from –50 to –20 °C, with an event at –30 °C; in Figure 4 crystallization is also delayed but to a abrupt change at –25 °C (A), and the event B during warming is very close to –21.7 °C. For pharmaceuticals the methods of freezing are limited: product in vials, seldom in trays, is frozen

on the shelves of the drying chamber, to reduce contamination risks; freezing by LN_2 is technically possible, but the expenditure is substantial. Product development should try to avoid this process. For food the freezing methods vary widely, as discussed in Section Food Plants.

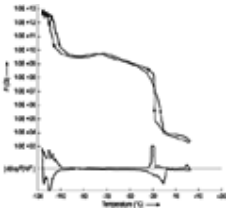

Figure 3. Electrical resistance as a function of temperature for a 1 % NaCl solution, cooling rate 3 °C/min, warming rate 3 °C/min, and the first derivative d(log*R*)/d*T* (Figure 1.22.a from [7]).

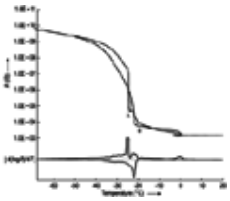

Figure 4. Typical electrical resistance plot of 1 % NaCl solution with 1 °C/min cooling rate, event A at –24.5 °C during cooling, event B at –21.8 °C during warming. (Figure 1.22.2b from [7]).

2.2. Main Drying

Sublimation of ice during main drying (MD) is mostly governed by heat transfer to the sublimation front of the ice with temperature T_{ice} and the transport of the water molecules from the ice front through the dried product is the second factor which influences the main drying time t_{MD} for product with layer thickness of less than ca. 2 cm in the range of 10 to 20 %. The main drying time t_{MD} can be estimated by Equation (2) [4]

$$t_{MD} = (\varrho_g \zeta_w (LS) \Delta m d) / T_{tot}\{(1 K_{tot}) + (d/2\lambda_g) + [d/2(LS)b/\mu]\} \quad (2)$$

where ϱ_g is the density of the frozen product (kg/m^3), ζ_w the fraction of water in the initial product (kg/kg), LS the sublimation energy (2.805 kJ/ kg), T_{tot} the temperature difference (T_{sh}–T_{ice}), K_{tot} the total heat transmission coefficient from the shelf to the sublimation front of the ice, λ_g the thermal conductivity of the frozen product, *d* the thickness of the layer (m), Δm the fraction of frozen water = 0.9 kg/kg, and *b*/ μ the permeability (kg h^{-1} m^{-1} mbar^{-1}) for water vapor through the dried product.

In this equation the following simplifications are made:

- The layer is infinite; energy is only transmitted from the shelf to one side of the layer.
- The vapor is only transported from the ice front through the porous dried layer.
- The frozen layer is not porous.
- Heat transport in the already dried layer is neglected.

The error resulting from the last assumption at T_{tot} = 100 °C is ca. 4 %, and at T_{tot} = 50 °C ca. 2 %.

Similar to the freezing time equation one can influence the drying time of a given product by two factors: the temperature difference T_{tot} (T_{sh}–T_{ice}) and K_{tot}, the total heat transmission coefficient from the shelf to the sublimation front of the ice.

To calculate T_{tot}, T_{ice} must be known. It cannot be measured directly by sensors (see Section Barometric Temperature Measurement (BTM)), but by barometric temperature measurement (BTM). T_{ice} is not only necessary for estimating t_{MD}; it is also the most critical parameter in MD. T_{ice} is the equilibrium result of the energy transported to the ice front and energy used to sublime the ice. The heat supply and the sublimation of the ice depend on two main factors: The pressure in the chamber (p_{ch} or p_c) and the temperature of the shelves T_{sh}; however, the structure of the ice, the pressure difference between the chamber and the condenser which transports the sublimed ice, and the number of vials in the charge have more or less measurable influences, T_{ice} also supplies certain information about the condition and the structure of the product. Figure 5 and its legend show: collapse of the product is clearly marked by T_{ice}; lowering of T_{ice} (from plot 2 to plot 3) reduces the mobility of the molecules also in the concentrated inclusions; freezing in LN_2 (plots 4 and 5) produces a more uniform mannitol product, which can be dried without the abrupt changes of T_{ice} in run 2 at the same temperature as in plot 2. A lower T_{ice} (4 °C lower, run 5) does not prolong the first 10 h of MD. The delay of ca. 1 h in the next 10 h is due to the delayed equilibrium after T_{sh} has been raised. Figure 6 shows an other method of T_{ice} analysis: all T_{ice} measurements of one run are analyzed for the frequency distribution. The data of Figure 6 show the result of 5 runs (b represents 2 runs). The conclusions are: freezing of egg albumin (a) on the shelves of the plant leads to a fairly uniform structure, the distribution of T_{ice} data is regular (more data might improve this picture), the standard deviation (sd) of the average (av) is small (ca. 0.1 °C); (b) freezing in LN_2 without thermal treatment (see Section Methods of Structure Analysis) results in a larger sd, a frequency curve somewhat extended to lower T_{ice} meaning some preference for small

crystals and some differences between two runs with the same data, the dotted plot having a larger fraction of small crystals.

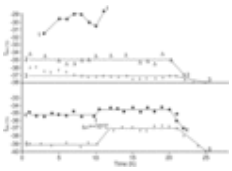

Figure 5. T_{ice} as indicator of the frozen structure of a 10 % mannitol solution, d = 10 mm. (Fig. 2.89 from [7]).

Runs 1 to 3: Product frozen in vials on the shelves to below –45 °C.

Run 1: sublimation rate too high for the possible water–vapor transportation, T_{ice} rises, product collapses, T_{ice} is irregular. Run 2: T_{sh} 0 °C, p_c 0.15 mbar, av T_{ice} –34.9 °C, sd 0.65 °C, product contains highly concentrated inclusions between ice crystals; depending on the position of the sublimation front, T_{ice} changes abruptly with the amount of concentrates being dried. Run 3: T_{sh} –5 °C, p_c 0.08 mbar, av T_{ice} –37.2 °C, sd 0.7 °C; due to lower T_{sh} and p_c the mobility of water molucules in the concentrate is reduced. In the first 6 h or so T_{ice} is a little higher than av; inclusions near the surface from freeze concentration are dried, after which the structure becomes more uniform, as indicated by a lower temperature of the ice. Runs 4 and 5: Product frozen in LN_2 and loaded on shelves at –45 °C, T_{sh} is raised at 10 h to +10 °C. Run 4: T_{sh} and p_c as in run 2, av T_{ice} –35.1 °C, sd 0.50 °C, after 11 h –34.5 °C, sd 0.6 °C. Run 5: T_{sh} and p_c as in run 3, av T_{ice} –39.1 °C, sd 0.22 °C, after 12 h –37.0 °C, sd 0.33 °C.

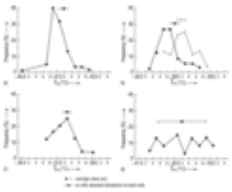

Figure 6. Frequency distribution of T_{ice} in 5 runs (2 runs together shown in b) (Fig. 1.78.2 from [7]. Data in Table 2.)

Table 2. Data of the runs in Figure 6

	EA*	EA	EA + 0.9 % NaCl	Pharmaceutical
Code in Fig. 6	(a)	(b)**	(c)	(d)
Cake thickness (mm)	20	20	20	20
Freezing method	shelf (sh)	LN_2	LN_2	shelf
T_{sh} (°C)	–5	20	20	0
P_c (mbar)	0.1	0.3	0.1	0.07
Average T_{ice} (°C)	–41.1	–29.2	–37.3	–42.2
sd (°C)	0.18	0.3	0.3	0.6

*Egg albumin.
**Two runs are shown in Fig. 6b with the same data, showing that LN_2 freezing without annealing can lead to slightly different structures and therefore slightly different T_{ice}.

The addition of NaCl and freezing in LN_2 (c) does not change the shape of the plot significantly compared with (b). The pharmaceutical product in (d), frozen on the shelf, does not have a uniform structure; if the dry product must have a more uniform structure, it should be frozen in LN_2.

2.3. Transition from Main to Secondary Drying (SD)

Temperature rise, pressure decrease, simultaneous measurement of pressure by thermoconductivity and capacitance gauges, and change in water vapor pressure by specific instruments are used to indicate the end of main drying. The many methods discussed show that all of them are not satisfactory to mark the end of the ice sublimation quantitatively.

As can be seen in Figure 5 and, in more pronounced form in Figure 7, average T_{ice} data are very constant as long as enough ice is available to fill the chamber with saturated water vapor after the valve between chamber and condenser has been closed; when the remaining amount of ice is unable to fulfil this condition, T_{ice} becomes systematically lower. The decrease in T_{ice} can be used to change the operating conditions from MD to SD, e.g., at maximum $T_{ice/n}$ = –1 °C or at maximum $T_{ice/n}$ = –3 °C (maximum $T_{ice/n}$ is defined as the maximum average T_{ice} of all T_{ice} measurements).

Figure 7. T_{ice} as a function of drying time. (O), (X): Measurements and repetition with the same product and the same process data; (•): Another product with other process data. 1) Average T_{ice}; 2) Standard deviation of T_{ice}; 3) Maximum $T_{ice/n}$ for o and × identical –22.53 °C; 4) Maximum $T_{ice/n}$ –1 °C, 4′) Maximum $T_{ice/n}$ –2 °C, 5) maximum $T_{ice/n}$ –3 °C. Maximum $T_{ice/n}$ = maximum of the average T_{ice} during MD (Fig. 6 from [5])

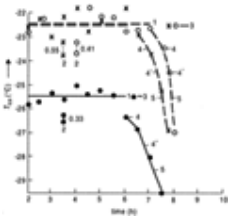

2.4. Secondary Drying (SD)

The desorption of the remaining water from the solids is called secondary drying. The moisture sorption behavior is often studied, e.g. by Fakes et al. [6] for mannitol, anhydrolactose, sucrose, D-(+)-trehalose, and dextran 40. The desorption process can be monitored by measuring the water desorption rate (DR), the desorbed amount of water in % of the solids of the product per hour (Equation 16a from [7])

$$\mathrm{DR} = 2.89 \times 10^2 \, (V_{\mathrm{ch}}/m_{\mathrm{fest}})(\mathrm{d}p/\mathrm{d}t) \tag{3}$$

where V_{ch} is the chamber volume (L), m_{fest} the mass of solids (g), dp the pressure increase (mbar), dt the measuring time for pressure increase (s). (By using 22.4×10^3 mbar L corresponding to 18 g H_2O, mbar L can be converted into grams).

Figure 8 shows the DR data of a 10 % sucrose solution with different number of vials, different freezing methods, and different T_{sh}. In all runs d = 10 mm and p_c = 0.15 mbar. The figure shows how the DR data sensitively reflect the various influences of freezing and MD process data: Run 1: T_{sh} too high, the product collapses; Run 6: longest t_{MD} (T_{sh}–T_{ice} = T_{tot} 7 % lower than in run 5), shortest t_{SD}, but dW = 1 % is practically reached in the same time as in run 5 (not shown in Fig. 8), since DR in run 6 during SD is larger than in run 5. Runs 3 and 4 are for LN_2 freezing, which has no influence on t_{MD}, but the desorption behavior is different, as plots 3 and 4 and the times for DR = 1 and 0.3 %/h show. Furthermore, one can see that the amount of uncrystallized water is larger for quick cooling: between DR 0.5 and 0.6 %/h the desorption is slower and the plot flatter. If the number of vials is increased and T_{tot} is decreased to 31.2 °C, as in run 2, t_{MD} increases, in agreement with Equation (2).

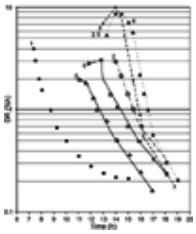

Figure 8. DR as a function of time with various process parameters (Fig. 2.91 from [7].)

Plot no.	Vial no.	FM[a]	T_{sh},[b] °C	T_{ice}, °C	t_{MD},[c] h	DR,[d] %/h	dW,[e] %	$v_{Tsh}/20$,[f] °C/min	t_{SD},[g] h
1	300	W	+30	product collapsed					
6	400	W	0	–32.3	10.15	14.3	14.6	0.12	4.45
5	100	W	0	–34.5	9.15	12.5	15.0	0.10	5.85
3	100	LN_2	0	–33.8	9.1	15.9	15.6	0.06	6.5
4	100	LN_2	0	–34.1	9.3	16.4	16.1	0.07	6.8
2	300	W	–5	–33.8	11.1	15.6	18.0	0.10	6.9

[a] FM: freezing method: W: vials charged on shelves at room temperature and cooled with the shelves, LN_2: product frozen in vials cooled by LN_2.
[b] T_{sh} = shelf temperature during main drying.
[c] t_{MD} = main drying time.
[d] DR (%/h): time in hours in which 1.0 %/h is reached.
[e] dW (%): time (h) in which the desorbable water content reaches 1 %.
[f] v_{Tsh20} / rate of temperature increase to 20 °C.
[g] t_{SD} secondary drying time (dW = 1 %).

If the DR data shown in Figure 9 are integrated over time the residual moisture content can be calculated in percentage of solids, as shown in Figure 10. The product is a 10 % albumin solution, filling height 10 mm, except for plots 5 and 6 (filling height 20 mm); T_{sh} in all plots 20 °C. In plots 1, 2, 3, and 5 the product was frozen on shelves, 4 and 6 in LN_2. Main drying in plot 1 at 0.5 mbar, all others at 0.3 mbar. In Figure 10 plots 2 and 3 are omitted for clarity. Runs 1 and 4 differ by the freezing method but mainly by the operating pressure p_c and therefore by T_{ice} –25.7 °C (1) and –28.8 °C (4). In spite of the smaller T_{tot} for run 1 (ca. 7 %), the pressure-dependent increase in K_{tot} (ca. 57 %) reduces the time to dW = 1 % from ca. 8.5h to ca. 5.7h (66 %). The pressure dependence below 0.1 mbar decreases.

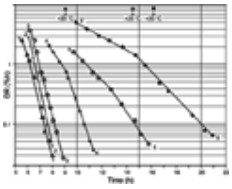

Figure 9. Desorption rates (DR) as a function of total drying time (Fig. 4 from [25]). 100 vials filled with 10 % albumin solution, with different freezing methods, different p_c, and different product thickness. Parameter: 1) 0.5 mbar (W), 2) 0.3 mbar (W), 3) 0.3 mbar (C) 4) 0.3 mbar (LN_2), 5) 0.3 mbar (W), 20 mm, 6) 0.3 mbar (LN_2), 20 mm.

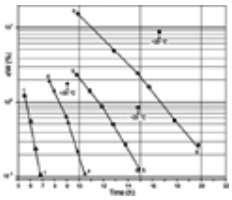

Figure 10. Residual moisture content dW as a function of time (Fig. 5 from [25]). 100 vials with 10 % albumin solution, with different p_c, different freezing methods, and layer thickness as parameter. 1) 0.5 mbar (W), 4) 0.3 mbar (LN_2), 5) 0.3 mbar (W), 20 mm, 6) 0.3 mbar (LN_2), 20 mm.

The DR and dW data monitor not only SD, but provide also information about the freezing rate, the amount of unfrozen water which could crystallize under certain conditions and which would otherwise delay reaching low dW values or make this practically impossible.

2.5. Packing and Storage

Vials as product containers are placed on the shelves with rubber stoppers already in a “holding” position, in which the water vapor can pass through canals in the stoppers from the inner of the vials to the drying chamber. The “packing” of the product takes place in steps: The stoppers can be pushed into the closed position under vacuum, or the chamber and therefore the vials are filled with a dry, sterile gas at a certain pressure, and then the vials are closed. The documented sterility of the gas is a technical problem (see Section For Pharmaceutical Products). The closing of several tens of thousands of vials requires many tons of pressure. The chemical engineering task with the stoppers is twofold:

1. During freezing the stoppers may shrink and slide into closed position, no drying is possible, and the product in such vials will melt. The canals must be large enough for water vapor flow without undue flow resistance, but they have also to be large enough to evacuate the vial parallel to the chamber, otherwise the overpressure in the vial will blow the stopper from the vial, the vial cannot be closed, and the stopper may impair the order of other vials and their closing. The stoppers and the vials have to be tested under these different conditions: on cold shelves, during evacuation (which speed is tolerable), and on heated shelves during SD.
2. The stoppers are steam-sterilized and absorb water during this process, which must be removed; otherwise, it will diffuse into the product during storage. During sterilization optimum stoppers take up ca. 1 % of water. Even with a stopper weight of only several grams, the amount of water can be 0.1 to 0.05 g. The amount of product in a vial will often be 1 g or less. 1 % residual moisture is then 0.01g. The equilibrium between water in the stopper and the product depends on the storage temperature and the sorption behavior of both. Fortunately, a certain amount of water in some stoppers is strongly bound to the stopper material (WANG et al. [8]) and can not desorb from the material at 100 °C, which is well above any storage temperature.

The time of safe storage of all pharmaceutical product depends on the storage temperature and the possible reactions of the product with water or within its own components. The water could defuse from the stopper to the product or be freed from the product if it is transformed from the amorphous into the crystal phase. This devitrification temperature T_g of glass-forming anhydrous carbohydrates varies from 120 °C for trehalose to, e.g., 77 °C for sucrose; it decreases with increasing water content of the dried product. On the other hand, RM can not be too low, since missing water molecules in a protein will cause its denaturation. The reactions between various components of the product or within one component do not stop at T_g; HATLEY and BLAIR [9] describe that this happens only at a temperature they call zero-mobility temperature T_0, which for trehalose is 44 °C and for sucrose 3.5 °C.

The packing and storage of food is governed by the large quantities to be handled (up to many tons of dry product per day) and the nonsterile conditions. The freeze-dried food is as hygroscopic as pharmaceutical products. Foods are freeze-dried in trays, which have to be vented in the plant with a very dry gas, often N_2, sometimes air, depending on the oxygen sensitivity of the product. Dry gas means a due point corresponding to the end pressure of drying, e.g., for 0.02 mbar, the due point of the gas should be ca. –55 °C. Once the pores of the dry product are filled with dry gas, some products can be handled for a short time in normal air, and often dry air is preferred. The container for long-term storage must be leak-tight (e.g., glass with a glued-on foil together with a screwed-on cover for protection or a can). A maximum water content of the product is required, this can be, e.g., 2–3 % or less than 1 %.

[Top of Page]

3. Temperatures and Pressures

3.1. Barometric Temperature Measurement (BTM)

During freezing temperature sensors can be used, since they are in relatively good and constant contact with the product. This is completely different during MD, since the ice front moves through the product and the product can contain ice or is already dry. The decisive factor is the temperature at the moving sublimation front, denoted T_{ice}. It can be measured by

the equilibrium water vapor pressure of the ice, the saturation vapor pressure p_S. If the valve between chamber and condenser is closed, the pressure in the chamber will rise to the equilibrium and this can be converted into the temperature of the ice (Table 3).

Table 3. Equilibrium water vapor pressure of ice (p_S) and related specific density of the vapor (ρ) (From [31])

t°C	p_S mbar	ρ_D g/m^3
– 70	$2.615 \cdot 10^{-3}$	$2.789 \cdot 10^{-3}$
– 69	3.032	3.218
– 68	3.511	3.708
– 67	4.060	4.267
– 66	4.688	4.903
– 65	5.406	5.627
– 64	6.225	6.449
– 63	7.159	7.381
– 62	8.223	8.438
– 61	9.432	9.633
– 60	$10.80 \cdot 10^{-3}$	10.98
– 59	12.36	12.51
– 58	14.13	14.23
– 57	16.12	16.16
– 56	18.38	18.34
– 55	20.92	20.78
– 54	23.80	23.53
– 53	27.03	26.60
– 52	30.67	30.05
– 51	34.76	33.90
– 50	39.35	38.21
– 49	44.49	43.01
– 48	50.26	48.37
– 47	56.71	54.33
– 46	63.93	60.98
– 45	71.98	68.36
– 44	80.97	76.56
– 43	90.98	85.65
– 42	102.1	95.70
– 41	$114.5 \cdot 10^{-3}$	106.9
– 40	0.1283	0.1192
– 39	0.1436	0.1329
– 38	0.1606	0.1480
– 37	0.1794	0.1646
– 36	0.2002	0.1829
– 35	0.2233	0.2032
– 34	0.2488	0.2254
– 33	0.2769	0.2498
– 32	0.3079	0.2767
– 31	0.3421	0.3061
– 30	0.3798	0.3385
– 29	0.4213	0.3739
– 28	0.4669	0.4127
– 27	0.5170	0.4551
– 26	0.5720	0.5015
– 25	0.6323	0.5521
– 24	0.6985	0.6075
– 23	0.7709	0.6678
– 22	0.8502	0.7336

– 21	0.9370	0.8053
– 20	1.032	0.8835

This pressure measurement should last < 3 s to avoid a temperature increase during the measurement time (sublimation decreases if the pressure approaches equilibrium, but heat transfer remains). To measure the increase in pressure accurately enough, the pressure should be recorded at a rate of 100 Hz. This can only be done by an automated system, and a computer can calculate the first derivative to determine the change in the slope of the pressure rise, which marks the equilibrium pressure, as shown in Figure 11. Besides the time limit two conditions for this measurement have to be fulfilled:

1. The amount of ice sublimed in the measuring time must exceed a certain ratio in relation to the chamber volume, since the accuracy of the pressure gauge has a certain limit (Eq. 4)

$$V_{ch}\mathrm{d}p/\mathrm{d}t < m_{H_2O}/t_{MD} \quad (4)$$

where V_{ch} is the chamber volume (L), dp = p_s–$p_{H_2O,ch}$ (mbar) the pressure rise during the measuring time, dt the time (s) until p_s is reached, m_{H_2O} the mass of water to be sublimed during MD, and t_{MD} the time of MD (s).

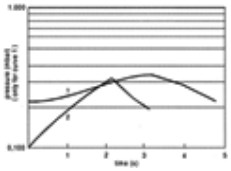

Figure 11. Pressure rise as a function of time. 1) Pressure rise in the chamber after the valve is closed; 2) First derivative of 1). The maximum of 2 is reached at 2.14 s, and the related equilibrium (saturation) vapor pressure is p_s = 0.286 mbar corresponding to T_{ice} = –32.7 °C. (Fig. 2 from [5].)

2. The leak rate of the chamber must be so small that the pressure increase due to this leak is small compared with the pressure rise due to the water vapor equilibrium pressure; e.g., in a chamber of 2000 L the pressure rise due to BTM is 4.3×10^{-3} mbar/s, corresponding to 8.6 mbar L/s; if the pressure rise due to the leak must be small (e.g. 1 %), the leak rate should not exceed 8.6×10^{-2} mbar L s^{-1}.

The pressure gauge in a freeze-drying plant should always be based on the capacitve (CA) and not on the thermal-conductivity (TM) principle. The thermal conductivity of a gas in a certain pressure range depends on the pressure but also on the gas type. In freeze-drying one can expect 90 % or more water vapor at the beginning of MD and less than 10 % water vapor at the end of SD. The reading of a TM calibrated in air at 0.1 mbar would be 0.05 mbar if the gas were pure water vapor. Furthermore the reproducibility of a TM is given by manufacturers as 20 % in the range between 10^{-3} and 10^{-2} mbar and as 15 % above 10^{-2} mbar. This leads to a calculated error in T_{ice} of ±1 °C, which in itself is too large, and the resuting uncertainty in the gas mix makes a TM unacceptable for freeze-drying.

The leak–rate measurement of a production plant should be done before each run. An undetected leak is a risk that can most likely not be repaired during a run and may ruin the total charge. The leak rate measurement and the hunting of leaks is described in section 2.2.9 of [7]. If the leak rate of a plant is measured at fixed time intervals, as described, it is possible to shorten the procedure before each run:

1. The history of the plant before the test must be standardized, e.g., the test is always done after sterilization of the plant, and its temperatures are all in the same range, e.g., 40 °C.
2. The evacuation time of the plant is recorded from atmospheric pressure to the operating pressure at the end of SD (e.g., 10^{-2} mbar). If this plot is within agreed tolerances at the end of the evacuation, one can assume that the plant is leak tight enough for the next run. Such a test should be done after sterilization since it insures not only sufficient leak tightness after the temperature and pressure changes, but also proves that no measurable amounts of water are hidden anywhere in the plant.

3.2. Pressure During Secondary Drying

The desorption of water from the solids depends for a given product only on the temperature, provided that the water vapor pressure in the chamber does not hinder the diffusion of the water vapor from the product into the chamber. The desorption isotherms describe the equilibrium between vapor pressure and the water content of the solid at a given temperature. In Figure 12 three desorption isotherms of a pharmaceutical product are given.

The water vapor pressure in the chamber should be small compared with the equilibrium vapor pressure, and the pressure of the permanent gases must be smaller than the vapor pressure in the chamber to facilitate the diffusion of the water vapor. The consequences for the pressure in the chamber (product as in Fig. 12) are, e.g., at an end temperature of 20 °C: to reach RM 0.2 % a water vapor pressure in the chamber $p_{H_2O,ch}$ = 2×10^{-3}mbar and for RM 2 % $p_{H_2O,ch}$ = 1×10^{-2}mbar are required; at 40 °C under the same pressure conditions RM 0.1 % and 0.8 %, respectively, can be reached. The final RM depends on the isotherms, which can only be measured with a certain effort in equipment and time. There is no relationship between p_c during MD and $p_{H_2O,ch}$ at the end of SD, but experience shows that the pressure during SD with T_{sh} in the range from 30 to 50 °C should be 10 % or more lower than p_c. Besides the pressure, the time to reach the desired temperature in the product depends on the thermal conductivity in the dried product, which can vary by 30–40 %, depending on the freezing method, for the same product (Fig. 8). For foods λ_{tr} can vary by a factor of 6. Since the desired

RM depends on the product temperature, the SD time depends not only on the pressure but also on the thermal conductivity of the dried product.

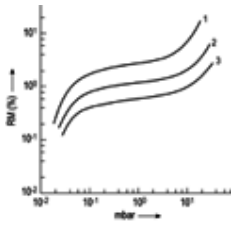

Figure 12. Desorption isotherms of the residual moisture content (wt %) as a function of water–vapor pressure. 1) 20 °C, 2) 40 °C, 3) 60 °C. (Fig. 1.85.6 from [7].)

Depending on the maximum tolerable product temperature and the desired RM for food, it can be possible that the required RM is reached without lowering the pressure for desorption or secondary drying. This means that temperature is high enough to reach the desired RM without lowering the pressure.

3.3. Water–Vapor Flow

During main drying the heat transfer to the sublimation front determines t_{MD}, but the plant must be designed to permit transport of the water vapor from the chamber to the condenser. The water vapor flow in the pressure range of freeze-drying, mainly from 0.8 to 0.01 mbar, is neither clearly viscous nor molecular; the mean free path of water vapor at 0.01 mbar and –40 °C is ca. 3.5 mm. The flow can be estimated from Fig. 13 (for details, see section 1.2.4 in [7]) which reflects some simplifications: The flow density is calculated for straight connections between chamber and condenser with the length to diameter ratio *l*/*d* as parameter. This does not account for a valve in the connection or for a bend. If chamber and condenser are flanged together and a mushroom valve is installed in the condenser, as in Fig 21, one can expect this configuration to have the same flow density as a connection with *l*/*d*≈1.6. A bend is characterized not by its geometrical dimensions, but by the “resistance” length, which is at least twice the actual length and can be five times as high; this depends strongly on the design and the pressure range of operation. For any production plant the flow density as a function of operating pressure is of vital interest, it should be measured once and than be used for the transfer of processes from a pilot plant to production. Example: In a pilot plant with *d* = 25 cm and *l*/*d* = 1.6 the vapor flow was 2.9 kg/h at 0.2 mbar and 0.37 kg/h at 0.06 mbar, showing that the plant is used at about 75 % water transport capacity (Fig. 13), and the flow density decreases more than linearly with pressure. The production plant has *d* = 80 cm; can it be operated under the same conditions with 25 times larger load than the pilot plant? At full flow capacity and *l*/*d* = 1.6 one can expect 40 kg/h at 0.2 mbar and 5 kg/h at 0.06 mbar, which allows only a loading of 13.5 times that of the pilot plant; even a valve with a large diameter of 1.1 m would only allow ca. 20 times the loading of the pilot plant. If the production plant is designed differently and has *l*/*d* = 5.0 (which could happen with one 90° bend) the flow capacity is reduced for the 80 cm valve to ca. 29 kg/h at 0.2 mbar and 2.5 kg at 0.06 mbar. With a larger *l*/*d* the decrease with pressure is more pronounced than with a smaller *l*/*d*. The measurement of the water vapor flow density is straight forward: a weighed amount of distilled water is filled (filling height ca. 12 mm) into the maximum possible number of vials, the water is frozen on the shelves to the minimum possible T_{sh}, and main drying is started for a given set of operating data, e.g., p_c = 0.15 mbar and T_{sh} = 20 °C. Drying is terminated when ca. 50 % of the ice has sublimed, the plant is vented, and the amount of water sublimed is determined by weighing. The weight loss divided by the time span after T_{sh} has been reached and the time of venting divided by the free area of the connecting tube between chamber and condenser is the flow density. This measurement is repeated at least at one or, preferably, two different pressures, e.g., 0.08 and 0.04 mbar. The three (two) measurements should be on one of the plots in Fig. 13, and *l*/*d* can be determined. If the resulting plot is to the left of the shown plots, *l*/*d* is larger than 5; nevertheless it can be used as the characteristic flow density curve of this plant.

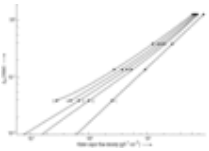

Figure 13. Density of water–vapor flow (g cm^{-2} h^{-1}) as a function of p_{ch} with jet flow (1) and *l*/*d* = 1 (2), 1.6 (3), 2.5 (4), and 5 (5) as parameter. 4) and 5) are not plotted below 4 10^{-2} mbar, as these data depend strongly on design details of the plant. They should be measured if needed. (Fig. 1.89 from [7].)

3.4. Residual Moisture Content

The drying time and the storage capability depend on the residual moisture content (RM) of the dry product. The methods for measuring this water content vary from chemical extraction and desorption under different conditions to analysis of the near-infrared spectrum.

The gravimetric method was obligatory according to the FDA until a few years ago: The weighed product (more than 100 mg) is stored in a gas tight container at 20 or 30 °C together with P_2O_5 and repeatedly weighed until the weight becomes constant. The weight loss is the RM of the product.

In thermogravimetric analysis the product (min. 2 mg) is placed in a microbalance in a temperature-controlled oven, and the weight loss is determined either at a constant temperature (e.g., 40 °C) or during a temperature profile (e.g., 5 °C/min from 30 to 100 °C). The weight loss is the RM of the product. To detect temperature-induced decomposition of the product, a mass spectrometer can be added to identify the molecular mass of the desorbed or decomposed molecules.

The Karl Fischer method extracts the water from the product into methanol and determines the amount of extracted water by titration with the Karl Fischer solution. In automated equipment the iodine for titration is produced by electrolysis and the

required amount of iodine for the titration is calculated by Coulomb's law. About 10 µg of water can be measured. The method does not work if the water cannot be extracted by methanol or the product reacts with iodine.

Water molecules have characteristic lines in the near-infrared (NIR) spectrum. The intensity of absorption in these lines is a measure for the amount of water molecules per unit volume. This method can be applied repeatedly in a vial without opening it. Since other components can have absorption lines overlapping with those of water, the measured spectrum must be mathematically analyzed; it varies with changes in the concentrations of the components. A calibration of the spectrum for water content is therefore only possible for one product. Within certain limits [10] the method can be used to check a large number of vials with product; the sample is not destroyed by the test. All methods depend on the binding energy between solid and water molecules. The results with each method can but need not be different from each other. RM data should be given with indication of the method of determination.

All methods described above can only be used after the drying process is terminated and the vials or the food products are removed from the freeze-drying plant. To determine RM data during SD can today only be done by the measurement of DR and the integration of DR over time resulting in dW; depending on the product temperature (e.g., 20 °C) only very weakly bonded water molecules will desorb, while at 40 °C additional molecules can be removed from the product.

3.5. Uniformity of Temperatures, Pressures, and Product

As noted in Section Chemical Engineering Tasks in Industrial Freeze-Drying uniformity of production is an important task for chemical engineering in freeze-drying. In a conventional production freeze-drying plant the reasons for a nonuniform quality of the product can be

1. Different flows of brine through each shelf (e.g., 12 shelves connected to one supply tube)
2. Different heat-transfer coefficients from the brine to the shelf within one shelf due to different brine flows in parts of the shelves
3. Delayed temperature increases or decreases due to different masses to be heated or cooled in parts of one shelf.
4. Certain areas for water–vapor transport are too small compared with the necessary flows in that area (to achieve the *l*/*d* ratio of 1.6 mentioned above, all other critical areas should be large compared with the area of the connection between chamber and condenser); or, differently stated: the pressure differences in the drying chamber have to be small compared with the operating pressure. This criterion becomes more important with low operating pressures, e.g., below 0.1 mbar and decisive below, e.g., 0.06 mbar
5. Influences of the door and wall temperatures on the product during freezing and freeze-drying.

The first point can be checked and corrected by conventional means. The second and third reasons are more difficult to quantify, since it must be done when the shelves are loaded. One possibility is to combine this test with the flow-density measurement described in Section Water–Vapor Flow. One shelf is selected as the model shelf. Around the entrance and the exit of the brine several temperature sensors are soldered to the shelf, as well as around the borders and in the center of the shelf. Furthermore, temperature sensors are distributed in the water in the vials in the same arrangement as those on the shelf. All sensor data are recorded during freezing and main drying. The sensors on the shelf around entrance and exit of the brine should show temperature differences during heating or cooling of 1–2 °C and less than 1 °C when the temperature is kept constant. Between the average of the inlet and outlet sensors the temperature difference can be 2–4 °C during heating or cooling and less than 0.5 °C when constant temperature is reached. The data given are examples of measured values; they may vary with the design, but the principle remains: a few °C are tolerable during temperature changes, but during constant-temperature conditions the temperature difference should be smaller than 0.5 °C. These measurements take some time and they must be taken once for a new design of a shelf. However, the user should ask the manufacturer for the results of such tests. Only then can the shelf design be excluded as a reason for nonuniform product quality. The temperature measurements in the water/ice during freezing will basically have a larger range of variations, since the freezing rate depends mainly on the degree of subcooling, the number of nuclei per unit, and the growth speed of the crystals. These conditions are not identical in all vials. But the average (av) freezing rate and its standard deviation (sd) between the regions on the shelf have to be the same: e.g., the av rate and its sd of all vials on the border of the shelf should be similar to those of all vials in the center or to all vials near the entrance and exit of the brine in the shelf. If the data on the shelf are satisfactory but those in the water/ice are not as acceptable it can be concluded that the influences of the door and wall temperatures disturb the uniformity. This can be confirmed if the sensors in the ice during MD show different av temperatures in vials at the border than in the center of the shelf. The door and wall temperatures can be substantially different, especially at low or high temperatures compared with room temperature, as shown in Figure 14. The consequences of these temperature differences on the water content of the product in vials (Fig. 15) are more drastic than expected. Without any shielding the product on the borders of the shelf shrinks, and RM is much too high; with the aluminum shields this product does not shrink, but the RM in the border vials varies widely, and only with a temperature-controlled shield do all vials have a uniform RM. The data are from a pilot plant (ca. 1000 vials); in a production plant with a greater mass of steel in the chamber and especially with a small amount of solids in each vial, the influence can be substantially more pronounced. It is a task of chemical engineering during pilot plant runs to define the necessary means of shielding a product and its process.

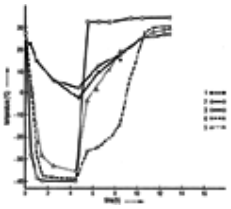

Figure 14. Temperature distribution during freezing (4.5 h) and freeze drying (8 h). Each shelf carries a frame for radiation shielding. The design can be different depending on the kind of vials and the loading and unloading system.

1) Internal temperature of the wall of the chamber, 2) Internal temperature of the door, 3) T_{ch}, 4) T_{pr}, 5) Temperature on the inner side of the radiation shield.

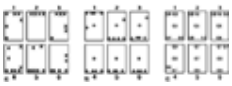

Figure 15. Residual moisture content in three identical runs with different shielding between vials and walls and door(s). (Fig. 1.68.2 from [7].) In each run six groups of vials where formed (168 vials each). Rows 4, 5, and 6 were close to the door, and rows 1, 2 and 3 close to the back wall.

a) No shielding, b) Shielding by an aluminum shield placed on the shelf, c) Shielding by temperature controlled shields

Residual moisture content distribution in the three runs

Run	a	b	c
Shrunk product			
rows 1–3	10 vials	0	0
rows 4–6	13 vials	0	0
RM in shrunk product	7/21 %	–	–
	av 15 %, sd 6.2 %		
Product in center vials	0.9/1.1 %	0.6/1.0 %	–
	av 1 %, sd 1 %	av 0.85, sd 0.15	
Product in rows 1–6	–	0.9/4.5 %	—
		av 1.9 %, sd 1.6 %	
All vials	–	–	0.4/1.0 %
			av 0.72 %, sd 0.15 %

For food products the temperature accuracy may be less stringent, but the temperatures of the frozen product can be as low as –40 °C and the shelf temperature during part of MD may be as high as 100 °C, resulting in a temperature difference between wall and shelf of ca. 80 °C. The shelves in a food plant can be formed in such a way that they are themselves the shields between wall and product. Such a solution is difficult in a pharmaceutical plant where easy cleaning, no corners, and rinsing of all water are “must” criteria.

[Top of Page]

4. Structures of Frozen and Dried Products

4.1. Influence of Freezing Method and Excipients

As shown in Section Freezing freezing rate, end temperature of freezing, excipients, and thermal treatment (see Section Methods of Structure Analysis) control the frozen structure. During drying this structure can only deteriorate by softening, collapse in micro-areas, collapse in total, modification of its amorphous structure, change from an amorphous structure to crystalline, growth in crystal size (recrystallization), and melting.

Most of these processes or events can be attributed to a certain temperature-time profile, a characteristic temperature, or a typical product behavior. In trying to generalize the influence of freezing, it becomes clear that many details contradict each other, such as fine crystal structure, all water crystallized, and no freeze-concentration of solids in water. There is only one general rule: If the freezing process and the events after the end of freezing and the beginning of drying are not well understood, the drying process can not be successfully engineered. An impressive example is given by BAKALTCHEVA and REID [11] with the drying of red blood cells (RBCs), a problem tried again and again over several decades. First one must find a cryogenic and drying protective agent. To stabilize proteins it is known, that sugar molecules can replace water molecules which are not removed during drying. Glycerol has been used to stabilize RBCs, but cannot be used in pharmaceuticals. The authors used adonitol, a polyol with similarities to glycerol, as protective agent in which they kept the cells for 12–14 h at 4 °C. Then they were frozen in a solution of 250 mM adonitol, 12 mM sodium phosphate dibasic, 2.9 mM sodium phosphate monobasic, and 20 wt % albumin, either on a precooled shelf at –30 °C with a cooling rate of 2.5 to 5 °C/min or in LN_2 at a rate of 75 °C/min. The LN_2-frozen product was kept for 48h at –30 °C. Both products were thawed for 10 min in a water bath at 37 °C. The survival rate after freezing on the shelf was 15 %, and after freezing in LN_2 80 %. From additional tests (not shown) the authors concluded:

1. For an optimum survival of RBCs a freezing rate of ca. 100 °C/min is necessary.
2. The damage to the LN_2-frozen cells is mainly a result of warming the cells from –196 °C to –30 °C (author’s note: crystals have time to grow and destroy the cells). The solution: Freezing in LN_2 down to only –30 °C and immediate freeze-drying at –30 °C (author’s note: no homogeneous crystallization of water at ca. –40 °C).

This example demonstrates the main problems of freeze-drying sensitive cells: Some water must be exchanged against a noncrystallizing agent, which at the same time delays the crystallization of the remaining water.

The freezing must be fast to attain the amorphous phase of water without crystallization or at least with only a limited amount of microcrystals. The sublimation time at –30 °C is not long enough to permit a change from the amorphous to the crystalline phase at the concentration of the excipients used. As shown in Figure 1 phase transition is a function of temperature, concentration, and time. Another example by MA et al. [12] shows the interdependence of various steps in the freezing and freeze-drying of tumor necrosis factor (TNF-MAb). The formulation consists mainly of 20 mg/mL TNF-

MAb, 10mg/mL maltose as stabilizing amorphous sugar, and 20 mg/mL glycin as bulking agent. The T_g of the product with in completely crystallized glycin was –42 °C. A process with heat treatment (see Section Methods of Structure Analysis) was developed: Cooling the product slowly (<1 °C/min) to –45 °C, slowly warming to –25 °C, and cooling the product back to –45 °C. The thus-frozen product could be dried at a T_{ice} of less than –25 °C (instead of below –42 °C), and in SD an RM of ca. 0.5 % could be reached in a normal desorption process, as confirmed by measurement of the desorption isotherm.

4.2. Methods of Structure Analysis

Structures of frozen and freeze-dried products are critical for the freeze drying process and the storage capability. Therefore, several methods have been developed to study them. Cryomicroscopy uses a cell for the sample which can be cooled, heated, and evacuated. The microscope is used to observe the frozen, the drying, and the dried product. Figure 16 was taken during quick cooling at ca. –25 °C, Figure 17 during the first warming process from –50 °C to ca. –35 °C and Figure 18 during the second cooling process at –55 °C. In Figure 16 the crystals (dark) are mostly uniformly distributed between the amorphous, concentrated solids (lighter color). In Figure 17 the crystals have grown, and some water from the concentrate has crystallized. In Figure 18 the boundaries between crystals and glass inclusions are more clearly visible, especially in the upper right corner; here the product was heat treated by warming to –35°C after the first quick cooling and brought back to –55 °C before drying; such a process is called thermal treatment (TT). A cryomicroscope permits a freezing and drying process to be followed visually. Crystal growth, softening of the structure, collapse, and melting can be recognized. The very thin layer is difficult to freeze slowly, and the behavior of the thin layer may be different from that of a 10 or 20 mm cake; nevertheless, cryomicroscopy is a very helpful tool for understanding the behavior of a product during freezing and freeze-drying. The scanning electron microscope is very effective for visualizing the structure of the dry product and its change during storage. Figure 19 shows the effect of freezing rate (a and b), collapse (c), and shrinking (d) on the structure of a product.

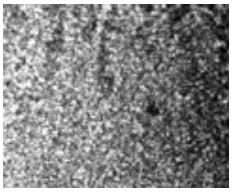

Figure 16. Cryomicrograph of a pharmaceutical product frozen at a rate of ca. 4 °C/min. During cooling at ca. –25 °C (Steris GmbH, Hürth, Germany)

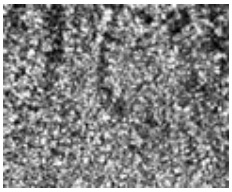

Figure 17. After cooling to –54 °C and thermal treatment (TT) at –36 °C (Steris GmbH, Hürth, Germany)

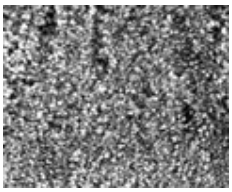

Figure 18. After cooling again to –55 °C (Steris GmbH, Hürth, Germany).

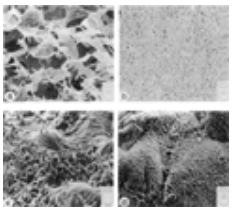

Figure 19. Photographs of a 1 % lactose solution by scanning electron microscopy. a) Frozen at 1 °C/min, b) Frozen at 150 °C/min, c) Collapsed dried trehalose solution from the bottom of the product, d) The same product shrunk after 6 months of storage at +20 °C with too high an RM and too high a storage temperature. (From [30].)

The other methods mainly used to study the freezing process are measurement of the electrical resistance (ER) as a function of temperature, thermomechanical analysis (TMA), differential scanning calorimetry (DSC), and nuclear magnetic resonance (NMR).

ER equipment is relatively inexpensive and straightforward to operate. Interpretation often requires experience with various products; its advantages are samples comparable with production freeze-drying and the possibility to simulate freezing on the shelf, as shown in Figure 20. In the lower part the 1st derivatve d(log$R$)/dT is shown, which greatly improves the visibility of events. In Figure 20a the sample vial with the product was immersed in the cooling medium, while in Figure 20b the cylindrical part of the sample vial was isolated by layers of plastic foils to simulate heat transfer from the shelf to the bottom of the vial. In Figure 20a the crystallization energy can be more quickly removed, part of the water crystallizes at around –10 °C, but another part solidifies without crystallization; this part softens during warming at event 1 and permits crystallization at event 2. In Figure 20b freezing takes place in two steps closer to equilibrium and events 1 and 2 can not be detected during rewarming. Conclusion: If this product is frozen on the shelf at 1 °C/min thermal

treatment to crystallize remaining water is not necessary. In thermal treatment or annealing after first freezing a product is heated to a temperature at which the water molecules can move enough to diffuse to existing crystals or nuclei; liquid water which may not have crystallized during the first cooling can crystallize during warming or the subsequent second cooling. Heat treatment can also be used to crystallize other ingredients which did not crystallize under the conditions of the first cooling. More details and applications are given in [7, section 1.1.5.1].

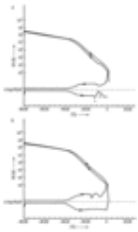

Figure 20. a) Electrical resistance of a pharmaceutical product as a function of temperature during cooling at 1 °C/min and rewarming at 3 °C/min. Heat-transfer medium surrounds the sample vial. b) Measurement of the electrical resistance as in (a), but the wall of the vial is insulated by a plastic tape up to the filling height of the product. Therefore, the heat is mostly removed through the bottom of the vial. (From [13].)

Differential scanning calorimetry (DSC) compares two different heat flows: one to or from the sample to be studied, and the other to or from a substance with no phase transitions in the range to be measured, e.g., glassmaking sand. The development of DSC equipment with a low heat capacity and quick response times has made it possible to modulate the normal temperature ramp by a sinusoidal temperature oscillation, a technique known as modulated DSC (MDSC). By Fourier transformation the total heat flow can be separated into reversing and nonreversing “kinetic” components. The reversing heat flow especially reveals glass transition events, which may be difficult to identify in the total heat flow signal or overlap with other nonreversing events. The reversing flow signal can be used to calculate heat capacity and its variation. The nonreversing component represents kinetic events such as crystallization and enthalpic relaxation; an example of MDSC is given in [14]. An other method to differentiate between reversible and the irreversible heat flow is the StepScan DSC [15]. The precondition for its use is a DSC system with a power compensation: sample and reference material are held in two separate calorimeters, each with its own heating elements. Both systems are maintained in equilibrium condition. The amount of energy to keep the equilibrium is directly proportional to the energy change in the sample. The power compensation system measures heat flows, whereas normal DSC systems measure temperatures. The StepScan method applies heating (e.g., 10°C/min) over a small temperature increment (e.g., 1.5–2 min), holds this temperature for a short time (e.g., 30 s), and calculates the heat capacity at that temperature. The irreversible or kinetic part of the total heat flow represents the “slow” processes which take place during the scan: enthalpic relaxation, crystallization, and melting. The two DSC methods permit the energy balances of the events to be quantified and the amount of product involved to be estimated. They measure changes in heat capacity (structural changes in amorphous products), and they can generally differentiate between irreversible processes (e.g., crystallization, melting) and reversible processes (e.g., glass transitions). The data obtained are quite accurate, but the equipment is much more expensive than an ER instrument; since the product quantities are very small, events in a product volume which may be several orders of magnitude larger and exposed to different heat-transfer conditions must be analyzed with regard to these two aspects.

Nuclear magnetic resonance (NMR) can be used to determine whether water is bound to other molecules (e.g., proteins) and therefore can not crystallize, how the collapse temperature T_c is influenced by unfrozen water, and the changes in a glass of highly concentrated solutions during warming from low temperatures below and above T_g. For example, YOSHIOKA et al. [16] studied the mobility of protons by NMR spectroscopy in freeze-dried bovine serum albumin (BSA) and gamma-globulin (BGG) and its relation with susceptibility to aggregation. The increase in spin–spin relaxation time t_{SR} of protons in BSA and BGG was strongly related to the water content in the range of 0.2 to 0.5 g water/g protein in both products and increased susceptibility to aggregation.

Thermomechanical analysis (TMA) has been used to measure the expension of the sample during freezing and warming. It has been helpful in explaining the breakage of vials during the warming of frozen mannitol solutions. Mannitol solutions above –25 °C expand ca. 30 times more than standard type 1 flint glass.

[Top of Page]

5. Industrial Freezing and Freeze-drying Installations

5.1. For Pharmaceutical Products

Sterility of the equipment and sterile handling of the product, which today is not sterilized as a finished product, are the compelling guidelines for the design of plants, the freezing and drying process, and packaging. Cleaning and sterilization of the plants is part of the total process and must be validated and documented during each run; both can only be fulfilled if the operations are mechanized and automated. They are therefore described in Section Process Automation. The freezing of pharmaceutical products in vials (or seldom in trays) is mostly done on the shelves of the plant. Freezing in LN_2 is technically possible and used, e.g., to freeze thick layers of product (35 mm) at high rates (> 10 °C/min). The mechanical engineering has to take care of the following special tasks in the design of freeze-drying plant: Water and condensate must freely flow from all parts of vessels and pipes; dead-end piping should have a minimum ratio of l/d = 2, and if possible 6; the door seal has to avoid condensate at the lower area of the door. The chambers are mostly cube-shaped with five critical parts: The door(s), the closing system for the stoppers, the shelves, the connection between chamber and condenser, and the venting system for the plant. The condenser is usually cylindrical and contains as main parts: the valve to shut off the chamber, the surfaces to condense the water vapor, the water drain, and the connection to the vacuum pumps. Figure 21 shows a schematic of a freeze-drying production plant with ca. 20 m^2 shelf area.

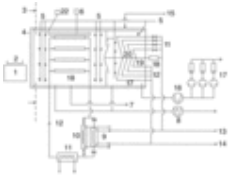

Figure 21. 1) Trolley for loading and unloading, 2) Vials in trays, 3) Sterile room, 4) Door, 5) Cleaning in place (CIP) system, 6) Pressure plate for closing stoppers with hydraulic system, 7) Water drain, 8) Water ring pump, 9) Heat exchanger for the brine, cooled by LN_2, 10) Heat exchanger for the brine, cooled by water (+20 °C), 11) Heat exchanger for the brine, electrically heated, 12) Brine (heat-transfer medium) circuit, 13) LN_2 inlet, 14) N_2 gas outlet, 15) Water for defrosting, 16) Roots vacuum pump, 17) Three two-stage pump sets, 18) Chamber, 19) Condenser, 20) Condenser plates, 21) Hydraulically operated valve between chamber and condenser, 22) Venting system

The chamber and condenser are in the same vacuum chamber, separated by a wall in which the valve is incorporated, providing the shortest possible path for the water vapor. The condenser and the brine heat exchanger are cooled by LN_2. The condenser surface consist of plates, and its temperature can be controlled between –110 and –60 °C. The shelves can be controlled by the circulated brine between –70 °C and +50 °C. The trays with vials can be loaded and unloaded from a trolley. The bottoms of the trays are removed before freezing. The shelves can be pressed together in one block and the trays are loaded onto the shelves by pushing one shelf after another in front of the trolley. Besides the shelf design (Section Uniformity of Temperatures, Pressures, and Product) and water-vapor transport (Section Water–Vapor Flow) chemical engineering is especially involved in the venting and the water–drainage systems, condenser design, and the optimum location of the vacuum pump connection. The gas-inlet filters (0.2 µm pore width) must be sterilized together with the plant. The integrity of the filters after sterilization must be proved and documented. The water intrusion test (WIT) is summarized: The known volume of the filter housing is flooded with sterile water at a defined pressure, and the change in pressure over time is measured. From this information the rate of water movement is determined. This is compared with the known data of an integer filter. After the test the filter and its housing have to be dried to remove all traces of water. Figure 22 shows the arrangement for the test. To completely drain the water after cleaning and sterilization, the chamber and condenser bottom should be inclined towards the drain, the water must flow freely from the vessels and through the piping, the valves have to be free of dead ends (diaphragm valves are preferred), and drain ports should have steam traps to prevent re-contamination from the drain lines. The condenser design is easy to formulate but becomes more and more critical to engineer with decreasing pressure: All water molecules must impinge on the condenser surface at least once, the partial pressure of the noncondensable gases must be small compared with the water vapor pressure at the entrance to the condenser, and the water vapor pressure at the entrance to the pumping system should be equal to the equilibrium pressure at the condenser temperature. The engineering aspects require effective cleaning and sterilization, a small temperature drop from the entrance to the exit of the refrigerant in the condenser coils or plates, and an acceptable temperature gradient from the ice surface to the metal surface of the condenser.

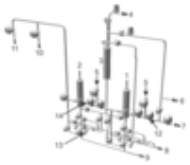

Figure 22. Filter integrity test by the water–intrusion method (WIT). (Fig. 2.46 from [7].)

1) Primary filter, 0.22 µm cartridge; 2) Secondary filter, 0.22 µm cartridge, backup filter can be sterilized independently from filter 1; 3) Test liquid reservoir (water for injection, WFI); 4) Inlet valve for WFI; 5) Pure steam inlet valves; 6) Inlet valve for filtered air (5 bar absolute); 7) Inlet valve for venting gas (1060 mbar); 8) Drain line with valves; 9) Connection for water ring pump; 10) Connection to condenser; 11) Connection to chamber; 12) Sanitary valves (diaphragm valves with no dead ends); 13) Temperature sensors

Figure 23 shows a schematic of a water vapor condenser. D is the mushroom-shaped formed valve plate, which directs the water vapor to the condenser surface. The condenser coils have a similar surface to the shelves for easy cleaning, and the coils are arranged, e.g., in three or four rows behind each other, as shown in Figure 24. The opening in one row is closed by the tube in the next row or in the next two rows. The distance between the two tubes is for the ice to build up, the next or the next two tubes in the next row or the two next rows have a similar distant from each other and from the other row(s) to provide space for the condensed ice. In this way the water molecules not condensed on the first row of tubes can condense on the tubes of the second or third row. To take care of the decreasing amount of vapor the condensers can have a cone for displacement of the vapor on the valve drive (Fig. 23) or a conical housing. The permanent gases are pumped off below the condenser coils as shown in Figure 23 and 25. If the gas is pumped off at a higher point, the condenser below this point will be gradually filled with gas (air is heavier than water vapor). This gas reduces the efficiency of condensation and finally stops it. If the gas pressure in the chamber or the condenser is not small compared with the vapor pressure, then instead of glassy ice, a porous snowlike ice is formed (Fig. 26), which can have ten times lower thermal conductivity. If the partial pressure of the permanent gases is too high it influences four steps of the process: It reduces water vapor transport from the chamber to the condenser (prolonging the MD time up to collapse of the product), it reduces the efficiency of vapor condensation even to the point of stopping vapor transport, decreases the thermal conductivity of the ice by a factor of up to 10, prolongs the SD time, and limits the possible minimum RM.

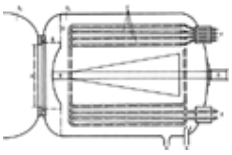

Figure 23. Schematic of a water–vapor condenser for a freeze drying plant (Fig. 2.18.1 from [7]). A) free diameter of the connection to the chamber, B) cylindrical opening by the movement of D, C) opening between condenser wall and valve plate, D) valve plate, hydraulic valve drive, E) condensation surface of the refrigerated coils, F) in- and outlet of the refrigerant, G) tube connection to the vacuum pump, H) water drain during defrosting of the condenser; p_{ch} and p_{co}: pressure in the chamber and in the condenser respectively.

Figure 24. Condenser with four rows of condenser coils. (Fig. 2.64.2 from [7]).

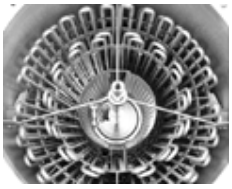

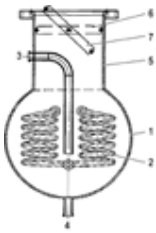

Figure 25. Schematic of a condenser design. (Fig. 1.90.3 from [7].) Condenser wall, 2) Condenser coil, 3) Suction tube of vacuum pump, 4) Water drain, 5) Connection to drying chamber, 6) Valve with seal in closed position, 7) Valve in open position.

This condenser design is not an optimum for three reasons: The open valve (7) restricts the vapor flow in the connection 5, water molecules can reach (4) without impinging (2), the vacuum connection 3 ends higher than the lowest condenser coil (2)

Figure 26. Coils in a condenser covered by ice, observed through a window during two freeze-drying processes. Left: smooth, solid surface. Right: porous, snowlike surface which typically occurs if the pressure of permanent gases during MD is too high (photographs by Otto Suwelack, 48727 Billerbeck, Germany).

The vacuum pumps have to be laid out for two criteria: Evacuation time of the plant to the operating pressure (recommended: 10 to 20 min) and capability to pump off the permanent gases during SD. Figure 27 shows the pumping speed of a three-stage combination: a Roots pump EH 500, a Roots pump EH 250, and a dry vacuum pump GV 80. The dry vacuum pumps operate without oil as a sealing medium, which avoids possible backstreaming of oil vapor into the condenser or the chamber. This pump set will evacuate a 4 m^3 plant (empty chamber and condenser) in ca. 16 min to ca. 10^{-3} mbar and has a capacity of ca. 300 m^3/h at 10^{-3} mbar. The gas desorption of the product and the plant is difficult to estimate at a pressure of 10^{-3} mbar, but to give an idea of the order of magnitude the following can be considered: The gas desorption of a product will be much smaller than the water desorption, which at the end of SD could be 0.1 %/h; at a load of 2000 g solids in the ca. 2 m^3 chamber this corresponds to 2 g/h, and degassing may be 0.02 to 0.2 g/h. The gas desorption of metals after 10 h of degassing is on the order of 2×10^{-9} mbar L s^{-1} cm^{-1} (chapter 2.4.1 in [17]). If condenser and chamber have a metal surface of ca. 50 m^2, they would degas 3–4 mbar L/h. The degassing of the product is estimated to be 0.02 to 0.2 g/h, which corresponds to 25–2250 mbar L h^{-1}. The pumping capacity (Fig. 27) at 10^{-3} mbar is 350 m^3 h^{-1}. Conclusions: the degassing of the metal surfaces in a freeze-drying plant can be expected to be small compared with the gas desorption of the product. The pumping speed as shown in Fig. 27 is sufficient to evacuate the plant in a reasonable time and to pump off the gas from the product. If the required end pressure of permanent gases is 10^{-2} mbar (example: see Section Pressure During Secondary Drying) the Roots pump EH 250 can be omitted.

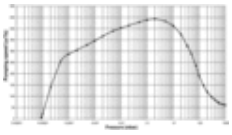

Figure 27. Pumping speed of a three-stage pump set: two dry roots blowers EH 500 plus EH 250, and one dry vacuum pump GV 80 with a high pumping speed from 0.001 to 10 mbar, end pressure ca. 10^{-4} mbar. (BOC Edwards GmbH, D-85551 Kirchheim/ München, Germany)

The automatic loading and unloading of production freeze-dryers becomes more and more the technique of choice; it is described in Section Process Automation.

5.2. Regulatory Requirements for Pharmaceutical Plants

The FDA defines (1993) the process validation as “establishing documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality attributes”.

The Guide to Inspections of Lyophilization of Parenterals, published by the FDA (July 1993) contains among others the chapters Lyophilization Cycle and Controls, Cycle Validation, and Lyophilizer Sterilization/Design.

In the European Union, the directive 91/356 EEC provides the principles and guidelines of Good Manufacturing Practice (GMP).

The following are a few of the most important European and U.S. guidelines [18]:

- GAMP: Good Automation Manufacturing Practice [19]
- EC-GMP Guidelines and Annex 15-PIC/S PI 006-1 and FDA Regulations (PIC/S = Pharmaceutical Inspection Convention) [20])
- User Requirements DIN 69905
- ISPE Commissioning and Qualification [21]

- The Parenteral Society — Specification and Validation of Freeze-dryers No. 9
- GXP: Good Practices for Computerized Systems in regulated GXP Environments [18]
- FDA: 21 CFR Part 211, § 211.68 (a, b)
- FDA: 21 CFE Part 11, “Electronic Signature, Electronic Records” [22]
- ISO: 100 11 Guidelines for Auditing Quality Systems [18]

The FDA defines a validation plan as a guideline of principles of general process validation. It should include the following:

1 Qualification and Validation Principles: This describes the principles of qualification and validation. It is a GMP requirement that manufacturers identify the validation work needed to prove control of the critical aspects of their particular operations.

2 Planning for Validation: All validation activities should be planned in a validation master plan (VMP). The plan should contain at least the following: validation policy, organizational structure of validation activities, summary of facility systems, equipment and processes to be validated, documentation format, planning and scheduling, change control, and reference to existing documents.

3 Documentation: A written protocol should be established that specifies how qualification and documentation will be conducted. The protocol should specify critical steps and acceptance criteria.
After completion of a satisfactory qualification, a formal release for the next step in qualification and validation should be made as a written authorization.

4a Design Qualification (DQ): The first element of a validation system or equipment should be the design qualification. The compliance of the design with GMP should be demonstrated and documented.

4b Installation Qualification (IQ): IQ should be performed on new or modified systems and equipment. IQ should include, but not be limited to, the following:
- Installation of the equipment, piping, services, and instrumentation, checked to current engineering drawing and specifications
- Collection of supplier operating and working instructions and maintenance requirements
- Calibration requirements
- Verification of materials of construction

4c Operational Qualification (OQ): OQ should follow IQ. OQ should include, but not be limited to the following:
- Tests that have been developed from knowledge of processes, systems, and equipment
- Tests to include a condition or a set of conditions and encompassing upper and lower operation limits, sometimes referred to as “worst-case” conditions
  The completion of a successful OQ should allow finalization of calibration, operating, cleaning and sterilization procedures, operator training, and preventative maintenance requirements. It should permit a formal “release” of the systems and equipment.

4d Performance Qualification (PQ): PQ should follow successful completion of IQ and OQ. PQ should include, but not be limited to, the following:
- Tests using production materials, qualified substitutes, or simulated products that have been developed from knowledge of the process and the systems or equipment
- Test to include a condition or set of conditions and encompassing upper and lower operating limits
- Although PQ is described as a separate activity, it may in some cases be appropriate to perform it in conjunction with OQ

5 Process Validation: This is the documented evidence that the process, operating within established parameters, can perform effectively and reproducibly produce a medical product which meets predetermined specifications and quality attributes.

6 Cleaning Validation: The cleaning process must be defined and the parameters described in writing. The limits and test procedures must be described.

7 Sterilization Validation: The sterilization process must be defined and the parameters described in writing. The limits and test procedures must be described.

8 Change Control: Written procedures should be in place to describe the actions to be taken if a change is proposed to a starting material, product component, process equipment, process environment, method of production or testing, or any other change that may affect product quality or reproducibility of the process.

9 Re-validation: This is a repeat of process validation to provide an ensurance that changes in the process/equipment introduced in accordance with change control procedures do not adversely affect process characteristics and product quality.

10 Risk Analysis: This is a method to access and characterize the critical parameters in the functionality of an equipment or a process.

11 Worst-Case Scenario: This is a condition or set of conditions encompassing maximum and minimum processing limits and circumstances within standard operating procedures which pose the greatest chance of a product or process failure when compared to ideal conditions.

Figure 28 shows the different steps in a qualification model following Good Automation Manufacturing Practice (GAMP).

Figure 28. Qualification model following GAMP guideline. (Fig. 6.3 from [7].)

The content of this section is not at all a complete description of the regulatory requirement for a freeze-drying plant to produce pharmaceutical products; it provides some examples of what must be considered in a long and complex process.

5.3. Food Plants

Food plants can be of the chamber or continuous type. Theoretically, food could be freeze-dried in chambers like pharmaceuticals, but the main problem is the freezing of products with different shapes and the handling of large amount of products; the cleaning requirements remain, but sterilization is not necessary. The freezing of food uses a variety of methods, of which the three main ones are: (1) freezing in a flow of cold gas, (2) freezing of droplets in a cold liquid and (3) prefreezing of a foam with final freezing in cold air.

Method 1 uses cold gas at a temperature from, e.g., –40 °C to the temperature of LN_2. If sticking together of, e.g., peas or leaves must be avoided, the surface of the product is quickly frozen with LN_2 in a CRUSToFREEZE plant (capacity 1500–5.000kg/h; Fig. 29) and finally frozen on a conveyor belt. In method 2 droplets of the liquid product are frozen in LN_2, as schematically shown in Figure 30. Coffee and tea extracts are foamed in method 3 by N_2 or CO_2 during cooling and partial freezing (e.g. to –5 °C) in a type of ice-cream machine. This foam must have a particular density, with the inclusion of certain amount of small ice crystals. The foam is cooled on a conveyor belt at –40 °C or below, as this product must pass a grinding and sieving system to achieve a desired grain size and density. The frozen product is dried mostly in two types of trays (Fig. 31). The ice sublimation rate of five types of trays is given in Figure 32. The tray types 3 to 5 in Figure 32 are more expensive than type 1, but the increase in the ice sublimation rate is 300 to 400 %. Food is mostly freeze-dried in heating systems installed in cylindrical chambers (Fig. 33) with outside (A) or built in condensers (B). The cylindrical chambers can be loaded and unloaded with product in trays: (A) by a carrier and transport device in which the trays (types 2 to 5 in Fig. 32) are loaded or (B) by an elevator which adjusts to the different heights between the heating plates and pushes the trays (type 1 in Fig. 32) on guide rails into and out of the chamber. In both cases, two to four of these tunnel sections can be put together, and vacuum locks be added on both ends. In design A a complete carrier with trays is moved into the lock, which is evacuated, and then the carrier is moved into the drying position, where the trays are lowered onto the heated shelves. In design B the trays remain on their guide rails and are pushed through the drying zone on one level, then an elevator brings them to the next level. The trays are brought individually through a lock into the chamber and through another lock back to the atmosphere. In the type-A plant up to 5000 t of freeze-dried coffee granules can be produced per year, and in the type-B plant up to 3000 t/a. Continuous freeze-drying plants have been built in which granules are transported on vibrating heated shelves. This method increases the water evaporation rate by a factor of ten or more. The limitation is the amount of product carried away by the dense water vapor flow or the need to filter the vapor flow with filters having a very large surface area to reduce the pressure drop through the filters.

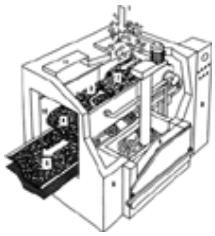

Figure 29. Schematic of the IQF freezing process CRUSToFREEZE (AGA AB, Frigoscandia Equipment AB, S-25109 Helsingborg, Sweden).

1) Product entrance, 2) IQF mixer, 3) Injection of LN_2, 4) Conveyor belt, 5) Belt for frozen product, 6) LN_2 pump, 7) Storage container, 8) Control unit, 9) Insulated housing, 10) and 11) Access openings.

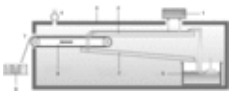

Figure 30. Schematic of a Cryogen Rapid Pelletizer. 1) Dropper unit, 2) Pelletizing channel, 3) Insulation, 4) Ventilator, 5) Product, 6) Conveyor belt, 7) Return channel for LN_2, 8) LN_2 pump, 9) Transport containers. (Messer Griesheim GmbH, D-47805 Krefeld, Germany.)

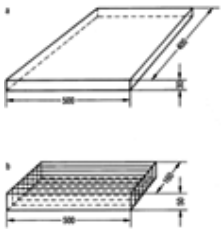

Figure 31. Types of tray for freeze drying cubes or granules of food.

Tray a: e.g., 400×500 mm, 30 mm height; tray b: e.g., 500×160 mm, 50 mm height.

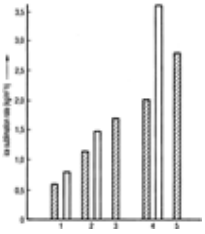

Figure 32. Ice sublimation rate (kg m^{-2} h^{-1}) for five different types of trays.

1) Flat tray as in Fig. 31a, with 20 mm filling height. 2) Flat tray with one rib of 20 mm height. 3) Ribbed tray, rib distance ca. 7 mm. 4) Ribbed tray with a V-shaped sieve between the ribs. 5) Ribbed tray for cubes or granules of food, designed by Otto Suwelack, D-48723 Billerbeck.

All tests were carried out with the same granulate. Hatched columns: heating plates or shelves were heated in ca. 30 min to 100 °C, and when the tray bottom reached +40 °C, the temperature was reduced in such a way that +40 °C was kept constant. White columns: heating plate or shelves were heated to +140 °C in ca. 2 h and the temperature reduced after +40 °C was reached as above (measurements by Otto Suwelack, D-48723 Billerbeck, Germany).

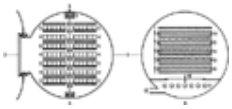

Figure 33. Schematic comparison of two commonly used tunnel freeze drying systems for the freeze drying of food.

1) Drying tunnel. In A: 2) Valve before the condenser, 3) Lifting and transport device, 4) Trays in transport position, 5) Trays on the shelves during drying, 6) Heated shelves. In B: 7) Radiation plates, 8) Trays, 9) Guide rails for the tray transport, 10) Separating plate between tunnel and condenser, 11) Condenser, 12) Vacuum connection.

5.4. Process Automation

5.4.1. *Prerequisites for Process and Plant Automation*

For the automation of freeze-drying processes and plants the qualities of the product, the qualities of the plants, and the developed process data have to be known, with emphasis on the acceptable tolerance of each datum and their cross dependence, e.g. p_c = 0.1±0.01 mbar may be tolerable if T_{sh} = 0±0.1 °C; If this tolerance cannot be achieved, and only 0±0.3 °C is possible, but at 0.3 °C p_c = 0.11 mbar is too high, than p_c should be specified as 0.09±0.01mbar. The plant data should include a failure analysis to highlight possible soft spots and eliminate them before automatization. The second requirement is to specify the measuring method for the data, e.g., all pressure gauges are of the capacitive type with temperature compensation for the range –40 to 60 °C, the reproducibility over a period of three months is better than ±0.003 mbar. To measure a temperature in the product during MD temperature sensors can not be used (the plant is automatically loaded and unloaded). During product development, at the latest in the pilot-plant phase, T_{ice} must be determined by BTM (see Section Barometric Temperature Measurement (BTM)), perhaps in parallel to temperature sensors in the product. Thermal conductivity gauges are too inaccurate for freeze-drying data and in addition depend on the type of gas, which varies in freeze-drying from 90 % or more water vapor to more than 90 % permanent gases desorbed from the product and the inner plant surfaces (see Section For Pharmaceutical Products). If different pressure gauges are used or T_{ice} is not measured, the whole product development and the pilot cycle may have to be repeated to gain documented data for an automation qualification and validation. These principles apply both for pharmaceuticals and food.

5.4.2. *Control of the Process and the Plant by Thermodynamic Data Measured During the Process: Thermodynamic Lyophilization Control*

The freezing of food is mostly automated due to the large volumes to be handled (see Section Food Plants). The loading and unloading of pharmaceutical freeze-drying plants is becoming more and more automated for two reasons: the handling of 10 000 to 100 000 vials is time consuming and burdened with the chances of several errors; the people handling the vials must sometimes be protected from the product, and each operator in the loading area increases the chances of product contamination. Two concepts of automatic loading and unloading are mostly used to day:

1. Automated guided vehicle (AGV): The vials are transported between the filling machine and the drying chamber or between the drying chamber and the capping machine by means of an AGV (Fig. 34).
2. Push and pull system: The vials are taken from the filling machine directly to the front of the Freeze-dryer by means of conveyors and pushed onto a shelf in the drying chamber (Fig. 35). Unloading of the drying chamber is effected over a pulling system: the vials are pulled out of the drying chamber and placed on a conveyor belt which transports them to the capping machine.

Figure 34. Sterile wall with two freeze drying plants (10 m^2 shelf area each), one loading storage (front) with ten divided shelves and one unloading storage (behind transfer cart). The loading and the unloading storage are served by the battery-driven transfer cart (in the background). (Novartis, Stein/Basel, Switzerland.)

Figure 35. Automatic loading and unloading system in front of the freeze drying plant (Steris GmbH, D-50354 Hürth, Germany.)

To separate product and operator completely the filling, transport and loading system can be enclosed in isolators as shown in Figure 36.

Figure 36. Automatic loading and unloading system completely enclosed in insulators. (Steris GmbH, D-50354 Hürth, Germany).

Cleaning and sterilization of pharmaceutical plants is becoming more and more automated for two reasons: human errors have to be reduced to a minimum and both processes have to be documented, qualified, and validated. Cleaning in place (CIP) has to take the four T principles into account [32]:

1. Temperature: Thermal effect; an increase in temperature of 1 °C increases the cleaning efficiency by 5 % (above 30 °C).
2. Turbulence: Mechanical effect; turbulence, as defined by the Reynolds number *Re* is basically a function of flow velocity and viscosity. For successful CIP the flow velocity in the pipes should be a minimum of 2 m/s (or *Re* > 8000); flat spray jets are used for cleaning the chamber, the pressure at the outlet should be in the range of 6 bar.
3. Time: Effect of the cleaning time.
4. Titration: Chemical effect; this is the influence of the chemicals in the detergents, cleaning agents or acids used. The concentration of these chemicals depends on the degree of equipment contamination.

A typical cleaning process consists of the following phases:

- Preliminary rinse with demineralized water (DE) or water for injection (WFI)
- Cleaning process with a cleaning solution or a detergent
- Final rinse, preferably with WFI

The process must be carried out in such a way that it can be validated. The following parameters are documented: water temperature, time, pressure, conductivity in the drain. CIP of Freeze-dryers requires extensive planning:

- Computer-controlled cleaning sequences with interlocks prevent unauthorized route-switching due to operator errors
- CIP processes can be sufficiently qualified and therefore fulfil CGMP requirements for validation
- The user develops a validation plan giving the highest and lowest limits for the verification of successful cleaning
- Verification of the correlation between the amount of product residuals in the last rinse water
- Optimization of the CIP process parameters until the desired level of cleanliness is reached

Steam sterilization is mostly used for freeze-driers. Only pure steam should be used; filtered plant steam has too high a content of permanent gases (regulation BS 3970). The recommendations of Section For Pharmaceutical Products should be applied for the plant design. The main design criteria are summarized:

- Free flow of the condensate from the vessels.
- Avoidance or minimation of “dead-ended” piping. The ratio of the pipe diameter to pipe length should be between 2 and a maximum of 6. (l/d = 2 to 6). Piping that comes into contact with operating media must also be sterilized. Such piping must be laid out with an inclination toward the chamber or condenser. Valves free of dead ends (sanitary valves) should be used exclusively. Diaphragm valves are recommended as the most appropriate valves.
- It must be possible to sterilize gauge heads. Particular care must be taken in choosing temperature-resistant gauges. A drifting of calibrated values at the high sterilization temperatures must be precluded.
- A certified sealing material which is FDA-compliant (21CFR 177.2600 and 21CFR 178.3570) should be used; e.g. PTFE or silicone.
- The temperature measuring points should be the coldest points in the system.
- Safety valves should be chosen which exclude any type of contamination of the vessels or piping when opened; in these valves the valve seat returns to the closed position after pressure has been released.
- If the door is equipped with a seal that provides tightness both in the vacuum and overpressure ranges, the door locking system must ensure that there is no gap between the door and the door flange at the lower areas of the door. Condensate can collect at the lower area of the door and chamber flange, and hence the sterilization temperature cannot be reached.

Steam sterilization can be replaced by the VHP (vaporized hydrogen peroxide) process, which works at ambient temperature and without pressure. Sterilization by VHP requires certain conditions which result from the nature of the H_2O_2 vapor:

- The plant must be dry, otherwise the H_2O_2 will dissolve in the water.
- The installation must be evacuated to ca. 1 mbar to ensure that the vapor reaches all parts and corners of the plant. H_2O_2 can still be applied at +5 °C, but at higher temperatures, e.g., +25 °C, the time to effect sterilization is shorter. Table 4 shows a typical set of *D* values (time in which the number of bacteria is reduced by a factor of ten) for *Bacillus stearothermophilus* for different surface temperatures and the necessary concentrations of H_2O_2.

Table 4. Typical *D* values (exposure time in which the number of bacteria is reduced by one order of

magnitude) for *Bacillus stearothermophilus* for different exposure times and the necessary H_2O_2 concentrations. (from [23])

Temperature of surface to be sterilized (°C)	Approx. H_2O_2 concentration (mg/L)	Typical *D* value (min)
4	0.3–0.5	8–12
25	1–2	1–2
37	3–4	0.5–1
55	10–12	0.02

The following technical prerequisites must be fulfilled:

Not all materials are resistant to H_2O_2. Stainless steel, aluminum, silicone, Viton, and Teflon are resistant; polyurethane, Plexiglas, and EPDM are not resistant in all compositions; and Nylon is not resistant. Pumps must be layed out to pump water which may be in the plant, but also to be resistant against H_2O_2. In smaller plants it is possible to use a catalyst to decompose the H_2O_2 in front of the oil-filled vacuum pump. In larger plants it is recommended to use a liquid-ring pump combined with an air injector for drying plant and venting the H_2O_2.

The sterilization process consists of three phases:

1. The installation is evacuated for drying to 1 mbar, or to at least below 10 mbar. For the documentation of the process the leak rate of the plant must be measured to prove that no outside air, perhaps contaminated, can enter the plant during phase 3.
2. For the sterilization phase a certain amount of H_2O_2 is injected into the plant and pumped off after some time. This procedure is performed once or may be repeated several times.
3. Removal of the H_2O_2: The vapor is pumped off by the vacuum pumps, the plant vented with sterile air up to 500 mbar, evacuated again, and so on. At the end of this procedure the residual H_2O_2 concentration is checked, e.g., by the Dräger H_2O_2 test tube [23], which indicates H_2O_2 concentrations down to 0.1 ppm.

The freezing of pharmaceutical products in vials by LN_2 can be documented by the temperature of the bath and the time involved. To document freezing on the shelves of the plant the inlet and outlet temperature of the brine in the shelves during freezing can be recorded, as in Figure 37, which permits the following conclusions from these runs:

– In this installation and with this load the beginning and the end of crystallization can be followed. At 20 min ca. 8 %, at 50 min ca. 60 %, and at 70 min ca. 90 % of the total energy has been removed from the product. The period 20 to 70 min is called here the main freezing time.

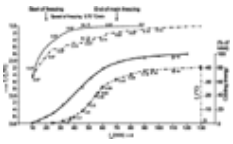

Figure 37. Thermodynamic data during freezing. All four curves are the average of two runs.

1) difference between inlet (T_i) and outlet (T_o) temperatures of the shelves of the plant with empty vials. 2) Same temperature difference during cooling with filled vials. The standard deviations between these 2 runs are given as figures on curves 1 and 2. 3) Energy used for cooling and freezing the product in percent of the total energy used. 4) Temperature in the product measured by sensors (2 sensors in 2 vials during each run). The curve is the average of these four temperatures, the numbers are the standard deviations of these 4 measurements. (Fig. 1 from [5].)

– The degree of subcooling was not uniform in all vials, as can be seen from the steps in plot 2. This is confirmed by the large standard deviations of the four sensors (plot 5): In the beginning between 30 and 60 min sd goes from 1.5 to 4.5 °C, then back to 0.5 °C, and in the end is 0.14 °C.

The optimum freezing time can be estimated by Equation (1) and compared with data as gained similar to Figure 37.

MD is controlled by T_{sh} and mainly by p_c (see Section Barometric Temperature Measurement (BTM)), as shown in Figure 38. T_{ice} is the temperature at which the heat input and the energy consumption by sublimation is in equilibrium. This temperature depends also on T_{sh}, the number of vials, the product structure and height, and the water–vapor flow density in the plant. (see Figs. 5-7). The plot in Figure 38 is only valuable if all these conditions are the same. T_{ice} follows a change of pressure in seconds, while a change in T_{sh} may take 30 min or more. For the control of T_{ice} the pressure is the most effective instrument. The optimum time of MD can be estimated by Equation (2). The quantitative change from MD to SD is described in Section Transition from Main to Secondary Drying (SD). The change can be automated by combining the steps of the change with decreases in T_{ice} (see Section Transition from Main to Secondary Drying (SD)), e.g., maximum $T_{ice/n}$ –1 °C, change T_{sh} from 0 °C to 30 °C; maximum $T_{ice/n}$ –2 °C, terminate pressure control p_c; limit the time span for the pressure to reach, e.g., 5×10^{-2} mbar to 15 min.

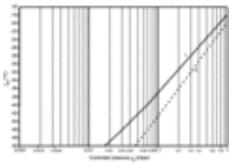

Figure 38. T_{ice} as a function of p_c. The shown plot is only valid under certain conditions.1) T_{ice} as a function of p_c, 2) saturation vapor pressure of ice. (Fig. 2.88 from [7])

SD is governed by three factors: the rate at which $T_{sh,SD}$ is reached, $T_{sh,SD}$ itself, and the pressure at which the desorption takes place. The course of SD can not be calculated. The optimum time in a plant is given (1) by the time for $T_{sh,SD}$ to reach its required level, e.g.. 30 °C at 2 °C/min, (2) the thermal conductivity of the product (the time for T_{pr} to reach, e.g., 28 °C is mostly the largest component of t_{SD}), (3) the water–vapor partial pressure must be small compared with the pressure given by the respective isotherm (example in Fig. 39), and the permanent-gas pressure must be small compared with the water vapor pressure. Automation during SD involves the repeated measurement of DR (e.g., every 15 min), the calculation of dW by the integration of DR over time, and automatic termination of the drying when the desired dW, e.g., 0.2 %, is reached. The dW data calculated from DR may not show the same figures as data determined by other methods, e.g., Karl Fischer method or thermogravimetry. Water in the dried product can be bound in many different forms, e.g., as surface water, as water bound more or less strongly to the dry substance, or as water of crystallization. Therefore, each method can lead to different results for different substances (for details, see [[7], sections 1.3.1 and 2.6.2.3]).

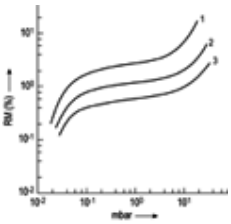

Figure 39. Desorption isotherms of the residual moisture content (wt %) as a function of water–vapor pressure. 1) 20 °C, 2) 40 °C, 3) 60 °C.

Thermodynamic lyophilization control permits, without temperature sensors in the product, automatically documented freezing, controlled and documented main drying, documented secondary drying, and termination of the total process at a required residual moisture content.

5.4.3. *Transfer of Pilot Plant Data to Production*

The transfer of pilot plant data to production often seems to be difficult, especially when in the pilot plant, e.g., 500 vials are loaded and the production plant handles, e.g., 50 000 vials. Five main reasons or a combination thereof impede the transfer:

1. The heat-transfer coefficients from the brine to the shelf surface are not comparable.
2. The water–vapor flow resistance from the chamber to the condenser surface as a function of pressure is different.
3. The influences of the wall and door temperatures on the product are not similar.
4. The condenser design and therefore its efficiency as a function of the operating pressure is not the same.
5. The pressure of the permanent gases during MD and SD is different.

For the following consideration it is assumed that methods and instruments of temperature and pressure measurements are identical, vials and the product are the same, and, very often important, the history of the product before the start of freezing and before the start of drying is the same.

One method to identify heat-transfer coefficients is described in Section Uniformity of Temperatures, Pressures, and Product. With shelf areas 100 times larger in production than in the pilot plant it is possible and even likely that the brine flow and the mass distribution of the steel in the shelves are not identical. One has to measure how important the difference is for the transfer of the process. During freezing the largest energy transport is required: During cooling and freezing one can expect ca. 400 kJ/kg of water, and during sublimation ca. 2800 kJ/kg. Cooling and freezing may take 0.5 to 1 h, the sublimation of the same product thickness takes 10 to 20 h; under these condition 800 kJ kg^{-1} h^{-1} during freezing and 280 kJ kg^{-1} h^{-1} during MD have to transported. It is therefore sufficient to compare the average temperature as a function of time in both plants on one shelf during freezing. Both shelves have to be loaded with the maximum number of identical vials filled with, e.g., 12 mm of water. The comparison of the average temperature is performed by a sufficient number of temperature sensors, e.g., 5 to 10 in the pilot and 20 in the production plant. The average difference between the plots of the two plants should not exceed 15 %, and the variation on the shelf in each plant should also not exceed 15 % from the average. If there are areas on the shelf with a deviation of more than 15 % the reasons for it have to be analyzed. 15 % is on the high side, 10 % should be more likely; therefore 15 % is mostly the starting point for critical differences.

Water-vapor flow is discussed in Section Water–Vapor Flow, including an example. It may be repeated here: With decreasing pressure (below 0.1 mbar) the possibility increases that the production plant does not have an opening between chamber and condenser comparable to that of the pilot plant. Before transferring a process to production, the water vapor flow density in both plants should be measured at two or three different pressures. These measurements in the production plant will not take longer than a few hours for the sublimation; the time for loading and unloading and weighing must be added. All can be done under sterile conditions if required.

The influence of door and wall temperatures on the uniformity of the residual moisture content is shown in Section Uniformity of Temperatures, Pressures, and Product. The guidelines for shielding have to be applied specially in large plants, since the mass of steel around the product will follow the shelf temperature more slowly. Even if a system is installed to cool the chamber more quickly after sterilization, it remains ineffective in adjusting the door and wall temperatures in time to the shelf temperature. There is no general rule for what type of shielding will be sufficient; it depends mainly on two factors: the larger the temperature difference between shelf and wall during freezing, main drying, and secondary drying, the more careful is the shielding required; the higher the operating pressure (e.g. < 0.1 mbar) the more the convection and the conductivity of the gas will contribute to heat transfer between walls and product; at pressures below 0.01mbar one may expect overwhelmingly radiation, which is easier to control than a mixture of all three contributions. Only during freezing on the shelves under atmospheric pressure can the influence of the three methods of energy transfer to the product not be avoided, except by using temperature-controlled shields.

The positions of pressure measurements in the chamber of the pilot and production plants must be evaluated. The pressure should be measured at a point of very low vapor velocity; at a point of high velocity the difference between static and dynamic pressure could theoretically reach almost a factor of two, but a difference of 20 % between pilot and production plant would lead, e.g., in the range of T_{ice} –30 °C to a difference of ca. 2 °C. This error would be five times larger than acceptable. High vapor speed will be near the connection between chamber and condenser. In the condenser the pressure should be measured in a part where the water vapor is in equilibrium with the condenser surface, where only this vapor pressure and the pressure of the remaining permanent gases is measured.

Differences in pressures between pilot and production plant during MD to obtain the same T_{ice} mostly indicate different flow resistance behavior of the two plants, which must be accepted unless the possible load of the production plant becomes too low. The other reason can be different leakage rates of the plants. Leak rates of both plants should be measured in such a case. (A leakage rate large enough to change T_{ice} measurably is unacceptable for DR measurements). A different pressure during SD at the same condenser temperature, leakage rate excluded, indicates a different pumping characteristic of the pump set. Some times the pumping speed at low pressures (e.g., below 5×10^{-2} mbar) in production plants is relatively lower than in a pilot plant. This can delay SD or make a small RM impossible.

[Top of Page]

6. Some Examples of Freeze-drying Processes

6.1. For Pharmaceuticals

Three processes for pharmaceutical products are described to demonstrate the reproducibility of a process the possibility of shortening the process time, and optimization of a process by the selection of excipients.

6.1.1. *Process Reproducibility*

In Figure 40 10 % sucrose solution is taken as a model substance to show the reproducibility of a total process. In the three runs the history of the product and the process data are kept as identical as technically possible. The standard deviation of T_{ice} in the three runs is 0.11 °C, and this indicates a high degree of uniform structure in the frozen product; T_{sh}, T_{co}, and p also show that the data were closely controlled. The result is shown in Figure 41 by the behavior of the three product runs during SD. The dotted lines represent the dW data calculated from DR measurements; the crosses, circles, and triangles each represent one measurement by thermogravimetry at 55 °C (see Section Residual Moisture Content). For these measurements vials were closed at the indicated time, and the residual moisture was determined after the end of drying. For the reproducibility of freezing and drying in this model of a pharmaceutical product one can conclude the following:

In spite of close freezing and drying control the three runs show small but measurable differences.

During main drying these differences are not measurable.

The time of change from MD to SD varies by less than ±0.5 h, depending on the decisions how to change (see Section Transition from Main to Secondary Drying (SD)), a variation of ±0.5 h can be expected and may be shortened by reducing the time between two measurements, e.g., 8 min instead of 15 min.

RG data (Fig. 41) before and during the change from MD to SD vary by up to a factor of eight. Thereafter the variation is more and more reduced and in the end (0.3 %) identical with dW.

The change from sublimation to desorption is not uniform in all vials; only when the residual moisture is on the order of 1 % of the solids does the whole charge become more uniform in this case.

The differences in the three dW plots result partially from point 3 above, but the differences in dW data show the necessity of SD documentation if small dW and/or small tolerances of dW are required.

If only 1 to 2 % dW is required ca. ±0.5 % from average can be expected.

Figure 40. Process data of three freeze drying runs, vials filled with a 10 % sucrose solution, filling height 10 mm.

1) Shelf temperature T_{sh}, 2) Product temperature by sensors, 3) Chamber pressure, 4) T_{ice} by BTM, 5) Condenser temperature. Double lines of the same data indicate the maximum and minimum values measured. T_{ice} for the three runs

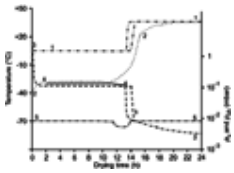

are: –34.99, –35.02, and –34.81, average T_{ice} of the three runs is –34.94 °C, sd 0.11 °C. Freezing rates: run 1: 0.7 °C/min, run 2: 0.65 °C/min, run 3: 0.75 °C/min. (Fig. 4 from [25]).

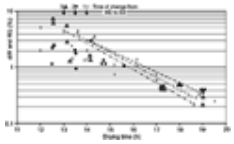

Figure 41. dW data calculated from DR measurement: runs 1, 2, and 3 dotted lines, compared with data from gravimetric measurements (RG): run 1 ×, run 2 •, run 3 ▲. For RG measurements 10 vials were closed at the shown time and measured after the end of SD. (Fig. 5 from [25]).

6.1.2. *Shorter Drying Cycle*

Figure 42 and 43 show the freeze-drying course of a Factor VIII solution. The product has a filling height of 27 mm and must be frozen in LN_2 at the rate of freezing that should be > 15 °C/min, otherwise the dry product becomes a powder if shaken. The ER measurement shown in Figure 44 requires $T_{ice} \approx$ –55 °C if frozen at 15 °C/min. If the product is thermally treated at –45 °C for ca. 9 h (the frozen product rests on the shelves at – 45 °C for 9 h and is then dried), the cryomicrographs (Fig. 45) show that the product can now be dried at $T_{ice} \approx$ –40 °C (the unfrozen water due to the high freezing rate has now being crystallized). From Figures 42 and 43 one can draw the following conclusions for the process:

1. T_{sh} can be raised in one step from –42 °C to e.g. –6 °C.
2. The pressure difference between the sublimation surface (–40 °C ≈ 0.128 mbar) and the chamber pressure (0.048 mbar) is almost a factor of three, which is unnecessarily large. In choosing a factor of two (which is conservative), p_c could be raised to 0.06 mbar, reducing the transported water-vapor volume by 25 % . Since the heat-transfer coefficient in this pressure range depends very little on the pressure, the higher pressure is only advantageous if the vapor transport from the chamber to the condenser is the bottleneck of the process (see Section Water–Vapor Flow). At a higher pressure, one could load more vials into the chamber. In the experiment described, vapor transport is not the bottleneck; therefore, one cannot expect a reduction in drying time as a result of higher pressure.
3. Based upon the earlier increase of T_{sh} (Section Barometric Temperature Measurement (BTM)) and the higher p_c (resulting in a slightly higher T_{ice}) the temperature increase to +20 or +40 °C, respectively, could be done earlier, probably by a few hours.
4. The temperature in the product rises only slowly, even after the increase in T_{sh}, (in 10 h from –20 to 0 °C and in a further 10 h from 0 to 30 °C). The temperature rise could be accelerated: The shelf temperature is increased, e.g., to 50 °C until the product has reached +25 °C. At that time T_{sh} is reduced to +40 °C. This temperature increase takes ca. 16 h with a temperature difference T_{sh}–$T_p \approx$ 30 to 35 °C. The higher difference of ca. 40 to 45 °C should again save several hours.
5. The pressure control could be cut off at 66 instead of 72 h, shortening the drying time by an additional 3–6 h.
6. In total, 10–15 h can be saved, resulting in a drying cycle of 64–59 h instead of 72 h, which could mean the completion of a cycle in 3 rather than 4 d. On the other hand, it is not likely that the drying time under the given conditions (filling height 27 mm and T_{ice} = –40 °C) can be reduced to 40 h.

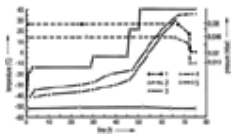

Figure 42. Freeze drying course of Factor VIII solution after freezing in LN_2.

1) p_{ch} (TM), 2) p_{ch} (CA), 3) T_{sh} , 4) T_{pr} (two sensors), 5) T_{co}.

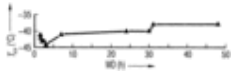

Figure 43. Plot of the temperature at the sublimation front (T_{ice}) of the run shown in Fig. 42 (Fig. 10b from [27]).

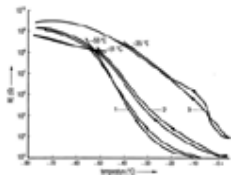

Figure 44. Electrical resistance as a function of temperature of three Factor VIII solutions with different excipients (Fig. 5 from [27].)

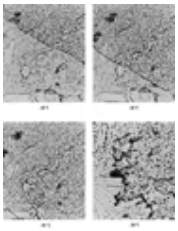

Figure 45. Photographs taken with a cryomicroscope of Factor VIII solutions at four temperatures. At –40 °C the structure is still visible, but is coarser than at –44 °C. At –35 °C the structure has collapsed. (Fig. 6 from [27].)

6.1.3. *Selection of Exipients*

The excipients are chosen for their cryoprotection during freezing and their lyoprotection during drying and storage. The used excipients also determine the possible T_{ice} during main drying and the storage temperature. Table 5 shows the $T_{g'}$ and UFW data of some exipients. $T_{g'}$ of a combination of exipients and drugs can not be calculated from the data of the pure substance; each combination must be measured. Figure 44 demonstrates the influence of different exipients on the structure of the product with the same drug (Factor VIII). The three products should be freeze-dried at T_{ice} = –55 °C (1), –51 °C (2), and –35 °C (3). The consequences of these requirements are shown in Table 6 and listed here:

1. The production of product 1 with the data as in Table 6 is practically impossible: In a plant having a valve between chamber and condenser with a diameter of 800 mm only 500 vials with 30 mL could be dried in one charge.
2. This picture is not substantially changed if product 2 is studied.
3. In both cases the following can be concluded: at T_{ice} = –45 °C p_c could be ca. 0.03 mbar and the load could be ca. 5000 vials. At –45 °C it is possible that the mobility of the molecules is small enough (the product behaves like a solid in the mechanical sense) to maintain a certain structure during MD. It is difficult to predict the outcome by analytical measurements. Tests in a pilot plant must prove an acceptable quality of the dried product. One should be aware of the fact that 1 or 2 °C more or less can change the outcome of such a test. It is therefore recommended to start the first test at the lowest possible p_c with a calculated number of vials. If this test is not sufficiently satisfactory between –45 and –48 °C the chances of producing an acceptable product with today's technnology are small.
4. At these low pressures p_c and low T_{ice} the required condenser temperatures are not the limitations, nor are the permanent-gas pressures.
5. The energy input and the permeability of the dry gas are also not the limit, as can be calculated by Equation (2): With T_{tot} = 30 °C and a heat-transfer coefficient K_{tot} = 33 kJ m^{-2} h^{-1} °C the main drying time would be estimated as ca. 60 h. K_{tot} = 33 is not the maximum possible, but at these low pressures only vials with a flatter bottom or smaller vials with more chances of contact with the shelf could reduce t_{MD}. In the end this might not even be helpful, since the limitation by the water–vapor flow remains.
6. The selection or combination of excipients can make the freeze-drying of a product impossible, technically difficult, or as normal as for many other products.

Table 5. $T_{g'}$ and UFW of some exipients (data from [26])

Excipient	$T_{g'}$ (°C)	UFW (%)
Sucrose	–32 (–33.6*)	35.9
Maltose	–30	20
Lactose	–28 (–29.7*)	40.8
Trehalose	–30	16.7
Fructose	–42	49.0
Glucose	–43	29.1
Glycerol	–65	45.9
Sorbitol	–43	18.7
Dextran	– 9	
Albumin	–10	
Poly(vinyl pyrrolidone)	–19.5 (–21.1*)	
Poly(ethylene glycol)	–13	
Sodium citrate/citric acid	–40	
Na_2PO_4/KH_2PO_4 (1/1)	–80	

*Data from 1998.

Table 6. Estimated production data for products 1, 2, and 3 in Figure 44

	T_{ice} (°C)	$T_{sh,MD}$ [a] (°C)	p_c [b] (mbar)	$T_{co,MD}$ [c] (°C)	$T_{sh,SD}$ (°C)	$T_{co,SD}$ [d] (°C)	p_{SD,H_2O} [e] (mbar)	$p_{SD,pg}$ [f] (mbar)	Load[g] (vials)
1	–55	–25/–6	$7\ 10^{-3}$	< –68	35	<–77	$1\ 10^{-3}$	$< 3\ 10^{-4}$	500
2	–51	–20/–6	0.015	<–62	35	<–63	$7\ 10^{-3}$	$< 2\ 10^{-3}$	1500
3	–35	–6	0.048	<–53	35	<–60	$1\ 10^{-2}$	$<3\ 10^{-3}$	10 000

[a] $T_{co,MD}$ is chosen similar as in product 3 (Fig. 44), resulting in T_{tot} (Eq. 2) ≈30 °C. For products 1 and 2 the difference between –6 °C and –55/51 °C would result in a temperature gradient over the ice thickness of 27mm at the beginning of the drying, which softens or collapses the product near the bottom of the vial. On the other hand –25 °C/–20 °C can not remain during MD since the dry product can not exceed –25/–20 °C, at which temperature the end of sublimation and the beginning of desorption (or the change from one type of drying to the other) may take days or more in this case.

[b] p_c in accordance with the data of Figure 38.

[c] $T_{co,MD}$ = maximum condenser temperature during MD.

[d] $T_{co,SD}$ = maximum condenser temperature during SD.

[e] p_{SD,H_2O} = water–vapor partial pressure during SD.

[f] $p_{SD,pg}$=partial pressure of permanent gases during SD.

[g] Load: approximate number of vials filled with 30 mL of solution (5 % solids) in a production plant having a valve diameter between chamber and condenser of 800 mm; the main drying time is estimated as 60h. The water–vapor flow density is taken from Figure 13 with *l*/*d* = 1.6.

6.2. For Food

Examples for freeze-drying food include mushrooms, which is relatively straightforward, and the freeze-drying of coffee extract, which has many more steps and is typical for the compromises in food freeze-drying: economy versus quality. In Section Food Plants plants are described with an output of 3000–5000 t/a of dry granules. Assuming a concentration in the coffee extract of 38 % solids, between ca. 8000 and 13000 t of concentrate must be treated annually, or ca. 1.5 t of extract must be treated per hour. The quality of the freeze-dried product starts with the quality of the raw beans, the roasting, and the extraction. Freeze-drying can preserve the qualities; it does not improve them. The freezing of the extract is described briefly in Section Food Plants. The goal of these steps is to produce frozen slabs, which can be broken and milled to granules of a desired grain size distribution and bulk density. These processes involve chemical engineering at low temperatures. The frozen, broken slabs have to be milled to, e.g., granules between 0.3 and 2.5 mm. The resulting powder, e.g., smaller than 0.1mm, can not be used for drying as it will close the channels between the larger grains and hinder or block the flow of water vapor, and the product melts and sinters together in areas with too much powder. The milling and sieving process must be performed below –40 °C; otherwise, the grains become too warm and stick together; the actual temperature of the air in those cold rooms is close to –50 °C. Figure 46 shows a compact granulating and sieving installation. Such installations are completely automatic; transportation of the granulates is achieved by vibration, but they have to be cleaned and maintained by personnel. Ice sublimed from the milled and transported product grows over time at the coldest places. Freezing and granulation of the extract is, as with many other food and nonfood products, the decisive phase of freeze-drying; the quality of the product produced here can be preserved not improved (but deteriorated). For drying, the granules are filled into trays (Fig. 31), each containing ca. 3 kg of frozen granules. About 500 of these trays must be automatically transported, filled, emptied, and cleaned per hour. The ribbed trays often have built-in sieves or other elements to transport the water vapor from the grains to the chamber; as shown in Fig. 32, only with such elements can high sublimation rates (2.5–3 kg of ice $m^{-2}\ h^{-1}$) be achieved. The main drying time with such trays is 7–8 h. Figure 47 shows a schematic of a continuous freeze-drying plant designed for ribbed trays. Four carriers for trays are used in each tunnel section. After the end of drying, gate valve 7 to lock 10 is opened and all carriers are moved one position forward. During this time the gate valve 7 to the entrance lock is closed, the locked door opened, and one carrier is moved into the lock. This is evacuated and the carrier is moved into the first position inside the drying tunnel, and so on. During movement of the carrier the trays are lifted from the shelves and placed back on the shelves when the transport of the carrier is finished.

The shelves in each tunnel section can be operated at a different temperature, in the first tunnel section the shelves can be operated at two different temperatures, permitting a temperature profile of six steps, e.g., 20 °C to sublime ice, which is condensed on the carriers and trays; 80 °C in the second half of tunnel section 1, tunnel sections 2, 3 and 4 40 °C and 30 °C in tunnel section 5 to allow the product to cool down to max. <40 °C before unloading. The condensers are not equally distributed but concentrated at the first two tunnels. Of the four condensers at the first two tunnels three are in operation and one is being defrosted; of the three condensers at the last two tunnels two are in operation, and one is being defrosted. If the plant is used for coffee granules the operating pressure can be between 0.2 and 0.6 mbar, depending on the grain size distribution and the type of trays used. Besides aroma, color, bulk density, and cost the stability of the grains against abrasion is a quality which must be achieved.

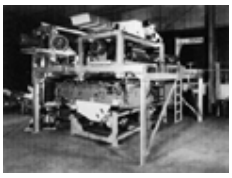

Figure 46. Granulating and sieving installation to produce a granulate of specified dimensions from frozen plates or large lumps. The mills can be seen in the background on top; the sieves can be seen in the lower part in front. The photograph was taken before installation in the cold room. (Photograph: Vibra Maschinenfabrik Schultheis GmbH & Co, D-6050 Offenbach am Main).

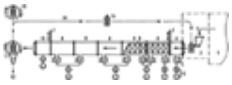

Figure 47. Schematic of a continuous quality controlled (CQC) freeze drying plant. 1) Preparation room, e.g., freezing, grinding, sieving, 2) Loading area, 3) Loading of trays and carriers, 4) Entrance lock, 5) Ice condenser, 6) Two-stage vacuum pump set, 7) Sliding gate valve between tunnel and lock, 8) Five drying tunnel sections, 9) Transport carrier with trays, 10) Exit lock, 11) Unloading of trays, 12 Exit of dry product, 13) Washing of carriers and trays, 14) Return of trays, 15) Carrier with empty trays.

6.3. For Metal Oxides, Ceramic Powders

In the publications studied it is difficult to distinguish between laboratory and small-scale production. Several papers concentrate on the production of small uniform particles which sinter at relatively low temperatures to densities close to that of the solid material. The term “spray freeze-drying” is used in one way or an other.

In a typical laboratory experiment ITATANI et al. [28] describe a process called “ultrasonic spray freeze-drying (USFD)” with two injecting nozzles. The two components - $Ca(CH_3COO)_2$/(hydrolyzed)$PO(OCH_3)$ (CaPo) and $ZrOCl_2/YCl_3$ (ZrY) are introduced by a flow of He into two ultrasonic systems. The resulting droplets are simultaneously frozen on a LN_2-cooled bottle. After freezing, the chamber is closed by a cover and evacuated by a rotary and oil-diffusion pumps. After 90 h of freeze-drying and some hot pressing the compact had a relative density of ca. 95 %, and grains smaller than 0.5µm were dispersed on boundaries of grains with sizes of 1–2 µm. Other methods of spraying are described, but the method is the same: the product is sprayed into fine particles and then dried. The formation of frozen particles and their drying are, in all cases studied, two separate steps.

[Top of Page]

7. Economic and Social Aspects, Outlook

Freeze-drying provides society with stable medical products for humans and animals, and it facilitates food and beverage preparation in specific areas such as spices, soup ingredients, and coffee. Many drugs would not be commercially available if they could not be stored under the conditions which prevail in many parts of the world. Freeze-drying of penicillin has made its application possible, freeze-drying of blood serum saved many lives in World War II, and its freeze-dryed fractions are still used to-day. Freeze-dried food is a more a matter of convenience and quality. Spices, e.g., can be air dried, coffee can be spray dried, soups can be canned, but the final products do not meet the original qualities or are heavy for transport. Economical freeze-drying needs substantial investments and crews of well-trained operators. In the medical field the cost to qualify processes and machinery add a significant part to the capital and labor cost. Quality and sterility have to be paid for. If the expenditure is not paid for by the markets, the product will not appear. The freeze-drying of food requires economically large installations with high throughputs; it is always a compromise between cost of production and required quality.

In the future of freeze-drying one can expect:

1. A substantial increase in the production of freeze-dried drugs
2. More sophisticated development and application of freeze-drying production processes
3. New food products will be available not by technical developments, but from marketing strategies of food manufactures

These expectations are based on the following current (2003) knowledge:

1. Worldwide more sophisticated freeze-drying production plants are being built. They replace older equipment without CIP and SIP or whose optimum operating pressures and/or freezing temperatures are too high. The new plants mostly have a substantially higher capacity per charge and are increasingly automatically loaded and unloaded. With lower temperatures during freezing and at the sublimation front of the ice (T_{ice}), the necessity of a uniform temperature distribution increases.
2. The development steps of a freeze-drying process are today often governed by trial and error; the transfer of process data to production is mostly based on incomplete data for the pilot and the production plants. In particular the influence of heat transfer and the flow of water vapor in both plants is not taken into account for process transfer. The process data will have to be measured and controlled by the actual data during the process and not by a predetermined time-temperature and time-pressure relations.
3. The qualification and validation of modernized or new processes and plants is time- and labor-intensive; therefore, points (1) and (2) will take time for their introduction; the effort is usually only justified for new products or production plants. On the other hand, the idea of a freeze-drying plant without low pressure (high vacuum), which has been raised again and again, will not become more feasible than it is today; the basic reasons for this are explained, e.g., in [7, section 1.2.6]. The other question —“why not make the process continuous”— can not be so easily dismissed (food is freeze-dried continuously). The main problems to be solved are sterility and abrasion during transport. Vacuum locks, different pressures during MD and SD, and condenser configurations could be possible. Products which could be sterilized after drying might be an answer to the first problem, if such products could be developed. Abrasion-free transport on air cushions requires a nonvacuum process; a magnetic field for support and transport seems to be conceivable, but the electrical energy to support it would require ca. 20 V in vacuum or installation outside the vacuum region. With all these “ifs” the question is “why”. The loading and unloading equipment and time can be dispensed with, but the drying time will not be shorter, and breakage of a vial may require stopping the process, emptying of the plant, and renewed start up, which alone takes the time of a full drying cycle. In food freeze-drying aluminum trays are used for bulk freeze-drying. Even if drugs could be sterilized after drying, the problem of dosage still remains. The end question may not be “why” but “who” is willing to invest the money for the development of a continuous plant for drugs, given the mentioned uncertainties and advantages that are difficult to

define.

4. Technical solutions for today's food freeze-drying of are many years old, sometimes one generation. They could be improved with modern technology. Only if new product ideas are developed and their marketing success is likely can one imagine the development of a new generation of food freeze-dryers. Otherwise it might be more economical to repeat proven plant designs with some engineering improvements.

The attempts to replace freeze-drying for certain drugs by other drying processes are sometimes described in the literature (estimated in 1 ‰ of all freeze-drying publications in the last few years). Solutions for replacement may be found (as for penicillin); newly developed drugs to be freeze-dried seem to outweigh those which can be dried with out freeze-drying. Overall an increase in the number of freeze-drying installations can be expected.

[Top of Page]

References

1. W. Umrath, *Kurzbeitrag für die Tagung Raster-Elektronenmikroskopie in Medizin und Biologie*, unpublished.
2. B. Luyet: "On Various Phase Transitions Occuring in Aqueous Solutions at Low Temperatures", *Ann. NY Acad. Sci.* (1960) 568. Links
3. G. Steinbach: "Die Bedeutung des Einfriervorganges, Berechnung des Gefriervorganges", VDI-Bildungswerk, RW 1570, Verein Deutscher Ingnieure, p. 3
4. G. Steinbach: "*Wärmeübertragung und Stofftransport bei der Gefriertrocknung. Berechnung von Gefriertrocknungsprozessen*", VDI-Bildungswerk, BW 1610, 1974.
5. P. Haseley, G. W. Oetjen: "*Equipment Data, Thermodynamic Measurements and In-Process Quality Control During Freeze-Drying*", PDA International Congress, Basel 1998, pp. 139 – 150.
6. M. G. Fakes, M. V. Dali, Th. A. Haby, K. R. Morris, S. A. Varia, A. T. M. Serajuddin: "Moisture Sorption Behavior of Selected Bulking Agents Used in Lyophilized Products", *PDA J. Pharm. Sci. Technol.* **54** (2000) no. 2, 144 – 149. Links
7. G. W. Oetjen, P. Haseley: *Freeze-Drying*, 2nd English Edition, Wiley-VCH, Weinheim, Germany, 2003.
8. Z. Wang, B. A. Frankel, W. Lambert, "Determination of Moisture in Rubber Stoppers: Effect of Karl Fischer Oven Temperatures", *PDA J. Pharm. Sci. Technol.* **55** (2001) no. 3, pp. 162 – 170. Links
9. R. H. M. Hatley, J. A. Blair: "Stabilization and Delivery of Labile Materials by Amorphous Carbohydrates and their Derivates", *J. Mol. Catal. B Enzym.* **7** (1999) 11 – 19.
10. P. Lin, C. C. Hsu: Determination of Residual Moisture in Lyophilized Protein Pharmaceuticals Using a Rapid and Non-invasive Method: Near Infrared Spectroscopy", *PDA J. Pharm. Sci. Technol.* **56** (2002) 196 – 205. Links
11. I. Bkaltcheva, T. Reid: "Lyopilization of Blood Cells", *International Conference on Lyophilization — Freeze Drying, Amsterdam, Oct. 2002*, International Society of Lyophilization-Freeze-Drying, Inc., Bala Cynwyd, PA 19004, USA, 2002.
12. X. Ma, D. Q. Wang, R. Bouffard, A. MacKenzie: "Characterization of Murine Monoclonal Antibody to Tumor Necrosis Factor (TNF-MAb) for Freeze-Drying Cycle Development", *Pharm. Res.* **18** (2001) no. 2, 196 – 202.
13. H. Willemer, unpublished results.
14. S. A. Knopp, S. Chongprasert, S. L. Nail, "The Relationship Between the TMDSC Curve of Frozen Sucrose Solutions and Collapse during Freeze-Drying", *J. Therm. Anal. Calorim.* **54** (1998) no. 2, p. 659 – 672. Links
15. W. J. Sichina: "Utilization de StepScan DSC pour l'optimisation des procedes de lyophilisation", *Spectral Analyse* **216** (2000) 35 – 37.
16. S. Yoshioka, Y. Aso, S. Kojima: Determination of Molecular Mobility of Lyophilized Bovine Serum Albumin and Gamma-Globulin by Solid-State ^{1}H NMR and Relation to Aggregation-Susceptibility", *Pharm. Res.* **13** (1996) No. 6, 926 – 930.
17. K. Diels, R. Jaeckel: *Vakuum Taschenbuch*, 2nd ed., Springer Verlag, Berlin, Göttingen, Heidelberg, 1962, pp. 238 – 239.
18. *GMP-/FDA-gerechte Validierung, Qualifizierung von Anlagen und Validierung von Prozessen und Systemen, Pharma-Technologie Journal*, Editor Cantor Verlag, Concept Heidelberg, D-69123 Heidelberg, Germany, 2002.
19. ISPE, *Good Automated Manufacturing Practice (GMAP) Guide for Validation of Automated Systems in Pharmaceutical Manufacture*, 2001.
20. PIC/S (Pharmaceutical Inspection Cooperation Scheme): Recommendation on Validation Master Plan, IQ, OQ, PQ, Cleaning (PI 006-1), 2001.
21. ISPE, *Baseline Pharmaceutical Engineering Guides, Vol.5, Commissioning and Qualification Guide*, 2001
22. *GMP-Gesetze der USA, 21 CFR 210/211 cGMP for Finished Pharmaceuticals*, 21 CFR 11 Electronic Records, Electronic Signature, Maas & Peither GMP-Verlag, Schopfheim, Germany, 2002.
23. R. Steiner:, "VHP Sterilisation of Freeze-Dryers", *ISPE-Seminar, International Society of Pharmaceutical Engineering (ISPE)*: Lyophilisation, Antwerp, Belgium, Nov. 1994.
24. Dräger: Prüfröhrchen for H_2O_2, Drägerwerk AG, Lübeck.
25. P. Haseley, G. W. Oetjen: *Oportunities and Limitations of Control of Freeze-Drying by Thermodynamic Data from the Process.*. Freeze-Drying of Pharmaceuticals and Biologicals. Center for Pharmaceutical Processing Research, in cooperation with Purdue University. A National Science Foundation Industry/University Cooperative Research Center, Vermont, 1998.
26. A. T. P. Skrabanja, A. L. J. de Meere, A. de Ruiter Rien, P. J. M. van der Oetelaar: "Lyophilization of Biotechnology Products", *PDA J. Pharm. Sci. Technol.* **48**, (1994) 311 – 317. Links

27. H. Willemer: “Freeze-Drying Process Data Determination for Human Blood Derivates with Factor VIII as Example”, *PDA Fourth International Congress*, Vienna, 1996, pp. 142 – 151.
28. K. Itatani, K. Akiyama, M. Aizawa, F. Scott Howell, I. Okada: “Sinterability of Apatite-Zirconia Composite Powder Prepared by Ultrasonic Spray Freeze-Drying Technique”, *J. Soc. Inorg. Mater. Jpn.* **286**, (2000) 194 – 202.
29. G. W. Oetjen: *Freeze-Drying*, Wiley-VCH, Weinheim, Germany, 1999, p. 6, Table 1.3.
30. P. J. Dawson, D. J. Hockley: “Scanning Microscopy (SEM) of Freeze-Dried Preparations: Relationship of Morphology to Freeze-Drying Parameters”, in J. C. May, F. Brown (Eds.): *Developments in Biological Standardization*, S. Karger AG, Basel, 1992, Vol. **74**, p. 185 – 192.
31. *Smithsonian Metrological Tables*, 6th ed., 1971 and VDI-Wasserdampftafeln, 6. Ausgabe, 1963.
32. F. Wolpers: *Clean-in-Place Systems for the Pharmaceutical Industry*, VDMA Seminar, Interphex 2001, Philadelphia.
33. A. P. Reinsert: “Factors Affecting the Erythrocyte During Rapid Freezing and Thawing”, *Ann. NY Acad. Sci.* **85** (1960) 576 – 594. Links
34. L. Riedel: “Enthalpy-Water Content Diagram for Lean Beef (also Valid for Other Meats with Fat Content Below 4 %)” in *Recommendations for the Processing and Handling of Frozen Foods*, 2nd ed.. International Institute of Refrigeration, Paris, 1972, pp. 28, 29.
35. G. W. Oetjen, H. J. Eilenberg: “*Heat Transfer during Freeze-Drying with Moved Particles*” International Institute of Refrigeration IIR (Comm. X, Lausanne 1969), pp. 19 – 35.
36. Z. Sharon, Z. Berk: “*Freeze-Drying of Tomato Juice and Concentrate, Studies of Heat and Mass Transfer*”, International Institute of Refrigeration IIR 132 (Comm. X, Lausanne, 1969), pp. 115 – 122.
37. O. M. Magnussen: “*Measurements of Heat and Mass Transfer during Freeze-Drying*”, International Institute of Refrigeration IIR (Comm. X, Lausanne, 1969), pp. 65 – 74.
38. R. D. Gunn, J. P. Clark, C. J. King: “*Mass Transport in Freeze-Drying, Basic Studies and Processing Implications*”, International Institute of Refrigeration IIR (Comm. X, Lausanne, 1969), pp. 79 – 98.

[Top of Page]

Abbreviations and symbols

av: average data

BTM: barometric temperature measurement

***bl* μ:** permeability of water vapor through the dried product, kg h^{-1} m^{-1} $mbar^{-1}$

CA: capacitive vacuum gauge

DR: desorption rate, desorbed water in % of solids per hour, %/h

DSC: differential scanning calorimetry

***d*:** thickness of layer, diameter, cm, m

d*p*: pressure rise, mbar

d*t*: time of d*p*, s, h

dW: desorbable water in % of solids, %

ER: electrical resistance, Ω

K_{su}**:** heat transfer coefficient from the heating medium to the freezing zone, kJ m^{-2} h^{-1} °C^{-1}

K_{tot}**:** heat transfer coefficient from the heating medium to the sublimation front, kJ m^{-2} h^{-1} °C^{-1}

LN_2: liquid nitrogen

LS: sublimation energy of water, kJ/kg

***l*:** length, m, cm, mm, μm, nm

λ_g**:** thermal conductivity of frozen product, kJ cm^{-1} s^{-1} °C^{-1}

λ_{tr}**:** thermal conductivity of dried product, kJ cm^{-1} s^{-1} °C^{-1}

MD: main drying, sublimation drying

m_{fest}**:** mass of solids, g

Δm: fraction of frozen water, kg/kg

NMR: nuclear magnetic resonance

p_c: controlled operating pressure during MD, mbar

p_{ch}: pressure in the drying chamber, mbar

p_{co}: pressure in the condenser, mbar

p_{MD}: pressure during MD, mbar

p_{pg}: pressure of permanent gases, mbar

p_s: saturation vapor pressure, mbar

RG: residual moisture content measured by gravimetric method, % of solids

RM: residual moisture content, % of solids

r: radius, m, cm

ϱ_g: density of frozen product, kg/m^3, g/cm^3

SD: secondary drying

SEM: scanning electron microscopy

sd: standard deviation

T_c: collapse temperature, °C

T_{co}: condenser temperature, °C

T_g: glass transition temperature, °C

$T_{g'}$: glass transition temperature if all freezable water is crystallized, °C

T_{tot}: time-weighted average of temperature difference (T_{sh}–T_{ice}) during MD, °C

T_{ice}: temperature of ice at the sublimation front, °C

$T_{ice/n}$: sum of all n T_{ice} measurements divided by n, °C

T_{MD}: temperature during MD, °C

T_{pr}: product temperature, °C

T_{sh}: shelf temperature, °C

TM: heat conductivity vacuum gauge

t_f: freezing time, h, min

t_{MD}: secondary drying time, h

t_{SD}: secondary drying time, h

UFW: “unfreezable water”, % of total water

V_{ch}: volume of a drying chamber, m^3, L

ζ_w: fraction of water in the initial product, kg/kg

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Good Manufacturing Practices

Standard Article

Derek Robinson[1]
[1]Monmouthshire, United Kingdom

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.h12_h02
Article Online Posting Date: January 15, 2004

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. GMP in the USA
3. GMP in the EU
4. GMP in the Wider World
5. The ICH Q7A Guideline
6. Quality Management
7. Principal Requirements of GMP
7.1. Personnel
7.2. Materials
7.3. Buildings, Facilities, and Process Equipment
7.4. Processes
7.5. Process Validation
8. Applicability of GMP
9. GMP in API Process Development

[Top of Page]

1. Introduction

Good Manufacturing Practices (GMPs) represent the minimum requirements that must be satisfied when producing pharmaceuticals for human or veterinary use [1]. They mandate a high degree of control over all stages of the manufacturing process, from ordering the raw materials through to packaging and dispatching the finished goods. All aspects of production must be described in written procedures, and all production activities must be thoroughly documented as they are carried out.

GMPs may be viewed as the third element of a trio of “GXP” regulatory requirements, which comprise also Good Laboratory Practice (GLP) and Good Clinical Practice (GCP). GLP mandates certain standards for laboratories engaged in scientific studies of new drug candidates, e.g., to obtain toxicological or pharmacological data from studies on animals. The GLP regulations enforced in the USA are detailed in part 58 of the Code of Federal Regulations Title 21. The equivalent requirements in the European Union are covered in Section II of the annex to Council Directive 87/18/EEC, amended by Commission Directive 1999/11/EC. GCP refers to basic principles in the conduct of clinical trials involving human patients. Within Title 21 of the U.S. Code, these issues are dealt with in parts 50, 54, and 56. The EU is currently harmonizing requirements for clinical trials across all its member states. A Clinical Trials Directive was published in the Official Journal of the European Communities in May 2001, and all states are expected to be in compliance with this by 2004. GMP governs the actual manufacturing of drug substances and products; it is mandatory for the production of clinical supplies of novel drugs and for commercial supplies of approved drugs.

This article will mainly focus on the GMP requirements at the stage of chemical manufacturing — the manufacturing of “drug substances” or “active pharmaceutical ingredients” (APIs). The requirements for “drug products” — i.e., the dosage forms such as tablets, capsules, injectable vials, etc — and for the manufacture of medical devices, are different and generally more stringent, but these are not discussed in detail here.

One of the principle reasons for GMP is to ensure the consistency of drug manufacturing processes. It avoids the situation, prevalent in other areas of the chemical industry, in which the final testing performed on individual batches or units is the

main, or even the sole, determinant of product quality. Reliance on testing alone is not appropriate for pharmaceuticals because:

Testing is only performed on a small sample, which may not be completely representative.
Defective products which escape detection at the testing stage constitute a grave danger to consumers (patients).
Patients themselves have no means of detecting such defects and cannot rectify them.

Testing cannot be relied upon to detect all failures in the manufacturing process; some defective batches will always get through. In some situations, such as the manufacture of socks or compact discs, this may not be a serious concern, since it causes only mild irritation to the 0.1 % (say) of customers who inadvertently purchase the defective objects, and they then have the option of returning them to the shop. With pharmaceuticals, the consequences of 0.1 % of patients receiving defective medicines are potentially far more serious. Therefore, GMP places as much emphasis on Quality Assurance — preventing the failures in the first place — as on Quality Control [2].

Even under the most tightly-regulated GMP regime, manufacturing problems and failures still occur, so Quality Control is still important. But under GMP, such failures should not simply be accepted as an unavoidable “act of God”. Rather they should be thoroughly investigated to determine the underlying causes and to prevent their recurring. By gradually eliminating the causes of failure, the incidence of failure should decline over time, and greater confidence in the consistency of the process is generated.

Under GMP, manufacturers are also expected to embrace novel technologies as they emerge, where these offer significant quality improvements. Examples are the use of more accurate engineering control systems, surface coating materials which are easier to clean and sanitize, and more sensitive techniques for chemical analysis. It is not that innovation must be embraced for its own sake; the cost of a novel measure must be considered in relation to the quality improvement it brings. However, the industry “benchmark” standards do evolve over time, hence prevalent use of the acronym cGMP (current GMP).

[Top of Page]

2. GMP in the USA

GMPs first emerged as a legal requirement in the USA in 1962, following substantial amendments to the 1938 Food, Drug, and Cosmetic Act, and since that time they have been enforced by the FDA (Food and Drug Administration) [3], [4]. The GMP regulations are detailed in part 211 of the Code of Federal Regulations Title 21. The current version has been in force since 1978, but FDA announced a major review of their GMP policy in 2002, which is expected to take about two years. The 21CFR211 regulations are legally binding requirements for all pharmaceutical manufacturers in the USA; drugs imported from other countries must also have been manufactured according to these standards.

Part 211 is divided into 11 subparts, covering the following:

A. General Provisions: Scope, Definitions.
B. Organization and Personnel: Responsibilities of quality control unit; Personnel qualifications; Personnel responsibilities; Consultants.
C. Buildings and Facilities: Design and construction features; Lighting; Ventilation, air filtration, air heating and cooling; Plumbing; Sewage and refuse; Washing and toilet facilities; Sanitation; Maintenance.
D. Equipment: Equipment design, size, and location; Equipment construction; Equipment cleaning and maintenance; Automatic, mechanical, and electronic equipment; Filters.
E. Control of Components and Drug Product Containers and Closures: General requirements; Receipt and storage of untested components, drug product containers, and closures; Testing and approval or rejection of components, drug product containers, and closures; Use of approved components, drug product containers, and closures; Retesting of approved components, drug product containers, and closures; Rejected components, drug product containers, and closures; Drug product containers and closures.
F. Production and Process Controls: Written procedures; Deviations; Charge-in of components; Calculation of yield; Equipment identification; Sampling and testing of in-process materials and drug products; Time limitations on production; Control of microbiological contamination; Reprocessing.
G. Packaging and Labeling Control: Materials examination and usage criteria; Labeling issuance; Packaging and labeling operations; Tamper-resistant packaging requirements for over-the-counter (OTC) human drug products; Drug product inspection; Expiration dating.
H. Holding and Distribution: Warehousing procedures; Distribution procedures.
I. Laboratory Controls: General requirements; Testing and release for distribution; Stability testing; Special testing requirements; Reserve samples; Laboratory animals; Penicillin contamination.
J. Records and Reports: General requirements; Equipment cleaning and use log; Component, drug product container, closure, and labeling records; Master production and control records; Batch production and control records; Production record review; Laboratory records; Distribution records; Complaint files.
K. Returned and Salvaged Drug Products: Returned drug products; Drug product salvaging.

In the mid 20th century FDA's enforcement activities concentrated mainly on secondary (dosage-form) manufacturing, and indeed the GMP regulations [5] were drafted entirely from the drug-product perspective. However, technically the U.S. legislation does not distinguish between drug products (dosage forms) and drug substances (APIs), and since the early 1980s FDA has increasingly emphasized the application of GMPs to API manufacturing also, using the 21CFR211 regulations as guidelines in their inspections of API plants. Initially, this was a source of much confusion and frustration among API manufacturers, because the regulations fail to take account of the very different manufacturing environment prevalent in API production plants, as opposed to pharmaceutical formulation plants. In practice, different inspectors had widely different expectations of how far the drug product regulations could reasonably be applied.

In 1987, FDA published their “Guide to Inspection of Bulk Pharmaceutical Chemicals” (revised 1991). But while this provided manufacturers with some useful insights into FDA's principal areas of concern, it is purely an internal agency document with no legal or regulatory status. Only in the late 1990s did official FDA Guidelines on APIs begin to emerge. Two draft versions were produced in 1996 and 1998, and the final version was endorsed in 2001. This final version is identical with the “Q7A Good Manufacturing Practice Guidance for Active Pharmaceutical Ingredients”, drafted by the International Conference on Harmonization (ICH), and thus is relevant not just in the USA but throughout the world [6].

FDA is required by its statutes to inspect all drug manufacturing facilities for GMP compliance once every two years — not just USA-based companies, but also foreign facilities where products are manufactured for export to the USA. Recently, however, the true inspection rate has declined to nearer once every six years, and the agency has been seeking ways to rationalize its inspectional work by concentrating on companies or areas of manufacturing which present the highest risks. Where an inspection reveals that a firm is not following the GMP regulations, its products may be deemed to be “adulterated”, regardless of whether or not they are physically contaminated or otherwise defective. Thus they may be barred from entry to the USA, seized, or even destroyed. In extreme cases, criminal penalties may be applied to the company and to individual company executives. In recent years fines of up to $500 million have been levied against several companies for GMP violations. However, most violative practices are associated with drug products rather than APIs.

[Top of Page]

3. GMP in the EU

In Europe, responsibility for drug regulation is split between various bodies: some national, some supranational. Until the 1990s, each country was essentially autonomous, maintaining its own agencies for drug approvals and inspections. For example, in the UK the Medicines Control Agency (MCA) fulfilled these functions. Each national agency operated according to its own procedures. The situation changed in 1995 with the establishment of the European Agency for the Evaluation of Medicinal Products (EMEA). EMEA now coordinates the approval process for new human and veterinary drugs throughout the EU, but most of the work involved is still carried out by the individual national agencies, now following centrally agreed procedures. In matters of GMP and inspections, a degree of international cooperation has existed for much longer, in the form of the Pharmaceutical Inspection Convention (PIC). Originally established in 1970 under the auspices of the European Free Trade Association (EFTA), it now comprises 24 (mostly European) countries who recognize the validity of each other's drug inspection procedures, with the aim of cutting down on unnecessary transnational inspections. Attempts have been made to agree mutual recognition procedures with the USA as well, but to date — in spite of a Mutual Recognition Agreement concluded in 1997 — these have been unsuccessful.

The European approach to GMP has been different from the American experience, in that there has been less emphasis on legislative enforcement. Rather, government inspectorates have sought to work in cooperation with manufacturers to encourage improvement in quality standards, in contrast to the adversarial and “investigative” approach associated with the FDA. The first GMP guide to be produced in the UK was published in 1971. Known as the “Orange Guide” after the color of its cover, this was originally intended as a Code of Practice for pharmaceutical manufacturers. Similar codes appeared in other European countries over the next few decades. From the late 1980s, the guidelines of the EEC, later the EU, were published, and eventually superceded those of individual member states. The Orange Guide continues to be published in the UK (latest edition 1997), but is now a straight reproduction of the European text [7].

The Orange Guide covers the same ground as the Part 211 regulations in the USA, but in a different order, and in somewhat greater detail. It is divided into nine chapters coving the following topics:

- One *Quality management*: Principle; Quality Assurance (QA); Good Manufacturing Practice (GMP) for medicinal products; Quality Control (QC).
- Two *Personnel*: Principle; General; Key personnel; Training; Personnel hygiene.
- Three *Premises and equipment*: Principle; Premises; General Production area; Storage areas; Quality Control (QC) areas; Ancillary areas; Equipment.
- Four *Documentation*: Principle; General; Documents required; Specifications; Starting and packaging materials; Intermediate and bulk products; Finished products; Manufacturing formula and processing instructions; Packaging instructions; Batch processing records; Batch packaging records; Procedures and records for Receipt, Sampling, Testing; Other.
- Five *Production*: Principle; General; Prevention of cross-contamination in production; Validation; Starting materials; Processing operations: intermediate and bulk products; Packaging materials; Packaging operations; Finished products; Rejected, recovered and returned materials.
- Six *Quality Control*: Principle; General; Good Quality Control Laboratory Practice; Documentation; Sampling; Testing.
- Seven *Contract manufacture and analysis*: Principle; General; The Contract Giver; The Contract Acceptor; The Contract.
- Eight *Complaints and product recall*: Principle; Complaints; Recalls.
- Nine *Self-inspection*: Principle

Additionally, there are 14 annexes, dealing with: Manufacture of sterile medicinal products, Manufacture of biological medicinal products for human use, Manufacture of radiopharmaceuticals, Manufacture of veterinary medicinal products other than immunologicals, Manufacture of immunological veterinary medicinal products, Manufacture of medicinal gases, Manufacture of herbal medicinal products, Sampling of starting and packaging materials, Manufacture of liquids, creams and ointments, Manufacture of pressurized metered dose aerosol preparations for inhalation, Computerized systems, Use of ionizing radiation in the manufacture of medicinal products, Manufacture of investigational medicinal products, Manufacture of products derived from human blood or human plasma.

GMP only became a legal requirement for drug product manufacturing throughout the EU in 1992, 30 years after its

imposition in the USA. At that time it did not strictly extend to API manufacturing, although API manufacturers were increasingly encouraged to follow GMP principles, and inspectorates took an increasing interest in their activities. Only with the ratification of the international Q7A guidelines in 2001 did APIs finally become officially regulated by European law.

[Top of Page]

4. GMP in the Wider World

The Q7A guideline is also official policy in Japan, the third main participant in the ICH process. Elsewhere in the world, most developed countries (e.g., Australia, Canada) have their own GMP regulations, but where these affect APIs they are rarely significantly different from Q7A. Developing countries have tended to adopt the guidelines from the WHO, which represent basic minimum standards for local pharmaceutical manufacturing, but are not considered suitable for the export of drugs elsewhere.

[Top of Page]

5. The ICH Q7A Guideline

The International Conference on Harmonization (ICH) is a joint forum between representatives of the regulatory authorities and the pharmaceutical industry from the USA, the EU, and Japan. It was established in 1990 with a remit to harmonize the technical standards for the registration of pharmaceuticals across the world. The main themes which it addresses are Quality, Safety, and Efficacy. Over the years it has developed many guidelines on matters such as impurities, residual solvents, clinical safety studies, stability testing, analytical validation, quality of biotechnological products, and much more. There has also been progress towards a “Common Technical Document” to serve as a new drug application format throughout the world.

Although GMP is not strictly a drug registration issue, and so is slightly outside its main remit, ICH became interested in this subject in the late 1990s. By this time, GMP standards for drug products were well advanced throughout the developed world, but were only enforced for APIs in the USA. FDA had issued their first “Guidance for Industry” on API manufacturing as a discussion draft in 1996, with a second draft appearing in 1998 after some consultation with industry representatives. Simultaneously, guidelines were being developed by the Pharmaceutical Inspections Convention (PIC) [8], the European Federation of Pharmaceutical Industries and Associations (EFPIA) [9] and the Pharmaceutical Research Manufacturers of America (PhRMA) [10]. ICH adopted these frameworks as the basis for their Q7A international guidelines, which first appeared in draft format in 2000, achieving final endorsement in 2001. Shortly thereafter, Q7A was officially adopted by the regulatory authorities in the USA, Europe, and Japan.

In the USA, Q7A has the status of a guideline, rather than a legally enforceable regulation. It represents FDA's current thinking on the subject and describes practices which will be acceptable to them. But conformance with the guideline recommendations is not mandatory; companies may adopt alternative approaches if they can be justified as consistent with the 21CFR211 regulations.

In any case, the guideline does not prescribe or recommend any specific measures, procedures, specifications, or controls; the emphasis is rather on ensuring that companies establish these for themselves. The word “appropriate” appears throughout, emphasizing that requirements and constraints may vary from one company to another. The guideline sets out the goals to be attained, but recognizes that they may be reached by different pathways. Each company is expected to define its own approaches and to demonstrate that these will succeed in meeting the prescribed goals.

The main topics covered by the Q7A Guideline are:

- Quality Management
- Personnel
- Buildings and Facilities
- Process Equipment
- Documentation and Records
- Materials Management
- Production and In-Process Controls
- Packing and Identification Labeling of APIs and Intermediates
- Storage and Distribution
- Laboratory Controls
- Validation
- Change Control
- Rejection and Re-Use of Materials
- Complaints and Recalls
- Contract Manufacturers and Laboratories
- Agents, Brokers, Traders, Distributors, Repackers and Relabelers.

There is also specific guidance for APIs manufactured by cell culture or fermentation and for APIs used in clinical trials.

[Top of Page]

6. Quality Management

For an organization to be compliant with cGMP, the most fundamental requirement is to have an effective quality management structure. Quality should be the responsibility of all persons involved in manufacturing, but the system must be defined and overseen by an independent Quality Unit, which is separate from all production activities. The manager of this Quality Unit should report to a senior level of the organizational hierarchy, and should not be subordinate to the production manager. The quality unit should be involved in all quality-related matters, and should review and approve all quality-related documents.

The quality unit should fulfill the functions of quality control (QC) and of quality assurance (QA), though these may be organized as separate units if desired – depending on the size and structure of the company. The QC function normally comprises the analytical work carried out to ensure that specifications are met — e.g. for raw material quality, equipment cleanliness or air quality. QA has a more wide-ranging role and has been defined (in Q7A) as “the sum total of the organized arrangements made with the object of ensuring that all APIs are of the quality required for their intended use and that quality systems are maintained”. The quality unit should have defined responsibilities, as detailed in the guidelines, including the responsibility to release or reject batches of the final APIs.

One important QU responsibility is to ensure that internal audits, or self-inspections, are regularly performed to determine whether established procedures are being followed and whether the cGMP regulations and guidelines are being observed. The audit findings and/or corrective actions should be documented and brought to the attention of responsible management.

Regular quality-reviews of APIs should be conducted at least annually, comprising in-process control and test results, batch failures, deviations, nonconformances, process or analytical changes, stability results, complaints, and returns. The results of these reviews should be evaluated and an assessment made as to whether any corrective action is required.

The emphasis that GMP places on Quality Management invites comparison with international quality standards such as ISO 9000, which offer much useful advice on the construction of a logical Quality System [11]. However, ISO 9000 is a generic standard, intended to be applied throughout all areas of industry and commerce, and does not specifically address the peculiar needs of the pharmaceutical industry. While it covers many of the same themes as GMP, such as the emphasis on management responsibility, resource management, product and service realization, measurement, analysis, and improvement, other important areas are missing, e.g., hygiene and sanitation requirements, which would not necessarily be relevant in other branches of industry. Under ISO 9000 the requirement is mainly to have written policies covering the relevant elements; it is left mainly to companies' own discretion what these policies should be. The GMP Guidelines tend to be much more prescriptive. Thus ISO 9000 registration may be a useful starting point for a company to achieve GMP status, but it is neither a necessary nor a sufficient condition.

[Top of Page]

7. Principal Requirements of GMP

The full text of the Q7A guideline is freely available from the FDA website [12], so only a brief summary is presented here. The main provisions can be divided into requirements for Personnel, Materials, Premises and Equipment, and Processes.

7.1. Personnel

Personnel are often the most significant source of variance in a process, and most failures can ultimately be traced to human errors. A drug manufacturer should therefore employ an adequate number of personnel with the necessary qualifications and practical experience to perform and/or supervise the required tasks. The structure of the company should be set out clearly in an organization chart, and all personnel should have specific duties recorded in written job descriptions. Training in the particular operations that the employee performs, and in the cGMP regulations that relate to his/her function, should be conducted regularly by qualified instructors. Records of such training should be maintained, and its effectiveness should be periodically assessed. Appropriate clean protective apparel should be provided, and staff should practice good sanitation and health habits.

7.2. Materials

All process materials must be carefully controlled to ensure they are only used in those processes for which they have been approved. This means maintaining proper storage facilities, where clear distinctions are made between materials which are quarantined, approved, or rejected/returned. Appropriate specifications should be developed for all starting materials, reagents, processing aids, solvents, water, and primary packaging materials, as well as for the final APIs and key intermediates. Each individual batch or lot should be uniquely identified and evaluated against its specifications before being approved for use in any process. The evaluation should normally be carried out by the quality unit; however, for some of the earlier less critical manufacturing steps, it may be performed by production personnel, following QU-approved procedures. All incoming materials should be sampled, and at least one identity test performed. In many cases it is acceptable to rely on a supplier's Certificate of Analysis as evidence of the required quality, as long as confidence in that supplier can be justified on the basis of a quality audit or other evaluation process and of satisfactory experience. Randomly selected lots of all incoming materials should be subjected to full specification testing from time to time. Whenever a material is used in a process, its lot number and the precise amount dispensed should be recorded on the batch production record, so that all materials are traceable back to their original sources. Records should also be kept on the disposition of each lot.

7.3. Buildings, Facilities, and Process Equipment

Buildings and facilities used in the manufacture of intermediates and APIs should be located, designed, and constructed to facilitate cleaning, maintenance, and appropriate operations, and to minimize the potential for contamination. There should be defined areas, or other control systems, for distinct operations such as receipt, sampling, quarantine, storage,

processing, packaging, and laboratory activities. Adequate lighting should be provided in all areas. Adequate ventilation, air filtration, and exhaust systems are required for those environments where the final API is isolated, dried, or packaged. Such areas should have smooth, continuous, and impermeable surfaces, with no sharp corners. API processes should ideally be physically segregated from one another; if this is not possible, then adequate procedures should be established which prevent cross-contamination.

Utilities such as steam, gas, compressed air, heating, ventilation, and air-conditioning should be qualified and appropriately monitored wherever they may affect product quality. Drawings for these systems should be available, and appropriate operating limits should be established.

All processing equipment should be of appropriate design and adequate size, and suitably located for its intended use, cleaning, sanitation, and maintenance. Any surfaces which directly contact the raw materials or products should be constructed of unreactive materials which will not adversely affect the quality of the API (e.g., steel equipment may not be appropriate for processing corrosive materials such as hydrochloride salts). All instruments should be regularly calibrated to ensure they consistently operate within defined tolerances.

Critical equipment and facilities should be formally qualified to demonstrate their suitability for the required operations and their conformance to predetermined specifications. Qualification will normally be conducted in different phases, beginning with a Design Qualification (DQ), in which the manufacturer's designs, blueprints, and specifications are compared with the user's own requirements to assess whether the basic concept will meet the defined process demands. Once the item has been installed and commissioned, it should go through the Installation Qualification (IQ) stage, which is a checklist to confirm that all the necessary functionality (switches, gauges, electrical connections, etc.) is present and of a suitable standard. The next stage is an Operational Qualification (OQ), in which the equipment is set to work on a typical task or process (usually employing inert materials), and details of its performance are collected and assessed. Finally, in the Performance Qualification (PQ), a trial run of a real process is conducted. (This PQ may be performed as part of the validation of that particular process.)

All equipment should be subject to a regime of planned preventative maintenance, to ensure that it continues to operate within its specifications, and to detect and rectify any potential problems at an early stage before they cause any batch failures.

In some cases it may be necessary to dedicate the equipment to particular processes, most notably for the production of highly sensitizing products such as penicillins.

Equipment should normally be cleaned to a satisfactory standard after each individual batch to prevent cross-contamination between batches and, most importantly, between different products. Equipment dedicated to one process or used for a multibatch campaign of the same product may be cleaned less frequently, but this should be justified on a case-by-case basis.

Companies are expected to establish their own cleanliness criteria based on an assessment of their own circumstances. Normally the criteria will include the requirement for no visible residue on the equipment surface. Where the surface is not completely accessible for visual inspection, some indirect means should be established, such as a rinse test or a swab test, with defined residue limits. Such limits should be rational, practical, achievable, and verifiable. For example, some companies operate on the 10 ppm principle, whereby limits are set to ensure that the following product cannot be contaminated with more than 10 ppm of the previous product. Other companies set limits to ensure that the maximum daily dose of the following product cannot be contaminated by more than 0.1 % of the minimum daily dose of the previous product.

Cleaning after each process should be defined in written procedures, which should be validated to ensure that they consistently achieve the defined standards. Cleaning validation involves identifying the areas of the equipment which are most difficult to clean and demonstrating that they are in fact cleaned to the required standard over a number of consecutive (usually three) runs. This may involve dismantling the equipment to some extent to reach otherwise inaccessible areas. However, cleaning validation need only be performed on a “worst-case” basis, that is, on those processes which are judged to present the greatest cleaning difficulties [13].

7.4. Processes

The production of each batch of API or intermediate should be closely documented to provide as complete a record as possible of the circumstances of its manufacture. To ensure uniformity from batch to batch, master production instructions should be prepared, dated, and signed by one person and independently checked, dated, and signed by another person in the quality unit. The master instructions should give details of all process materials required, including packaging and labeling, with their quality specifications and amounts. The major production equipment should be specified, along with detailed production instructions, including sequences, allowable ranges for process parameters, sampling instructions, in-process controls, time limits for completion of individual steps, and expected yield ranges at appropriate phases. Each individual batch should conform as far as possible to these master instructions, with any deviations being recorded, investigated, and evaluated.

Each individual batch production record should identify the operators and supervisors involved with the batch, record equipment numbers, the dates and times of each critical operation, the lot numbers of each ingredient used and the precise amounts dispensed, the actual results recorded for critical process parameters and in-process controls, actual yields at appropriate phases, details of any process deviations, and results of release testing.

7.5. Process Validation

A formal validation is required for all operations determined to be critical to the quality and purity of the API [14]. Process validation is often a very time consuming and laborious activity; therefore, many companies perform it only on the final

synthetic steps, relying on quality control and other GMP measures to assure the quality of the earlier, less critical intermediates. A written validation protocol should be established that specifies how the validation of a particular process will be conducted. This should identify the critical product attributes of the API (e.g., analytical specifications, physical form), and the critical process parameters (e.g., temperatures, stoichiometry) which may affect these attributes. For each critical parameter a normal operating range, a proven acceptable range, and, where appropriate, edge of failure limits, should be established. A series of consecutive validation runs (usually three) is then undertaken to establish that the process consistently falls within the defined ranges for each parameter and that the product consistently meets its quality specifications.

Validation studies should normally be completed before any commercial distribution of the API (Prospective Validation). For most new drugs, FDA approval will be withheld until the agency is satisfied that an adequate validation has been performed. In a minority of cases, an alternative approach (Concurrent Validation) may be more appropriate. For example, if the drug already has marketing authorization, its sales may continue while the validation studies are underway. (However, since validation has been a requirement since the late 1980s, there can be few drugs remaining in this category.) Another case would be where batches are made very infrequently because of low demand; in this situation the first batch may be approved and distributed before the validation is complete, the validation being resumed with the next batch. Concurrent Validation may also be used to justify approval of a single batch where there has been a substantial deviation from the established (validated) process. This would comprise detailed assessments of the impact of the deviation coupled with extra analytical investigations to ensure the quality of the drug substance. A third approach (Retrospective Validation) is mentioned in FDA guidelines, but increasingly discouraged. This involves no validation runs as such, but rather a detailed review of a large number (20–30) of consecutive historical batches, including any which have failed. This approach presupposes that sufficient details were collected at the time.

Once a process has been validated, strict controls must be enforced to prevent unauthorized changes. Any proposed changes to the process should be formally evaluated for the effect they may have on the critical quality attributes of the API. If a process change is judged to be major, i.e., it has the potential to influence the quality of the product, then its introduction should be preceded by a revalidation of the process. Minor changes which are judged unlikely to affect product quality may be implemented without formal revalidation, provided they give an equivalent input to the process. (For example, if the starting material is obtained from a new supplier, it should meet the same qualitative and quantitative specifications as that from the previous supplier.) Each process should be reviewed on a regular (perhaps annual) basis to assess how far it has shifted since its last validation, and a judgment should be made as to whether revalidation is required.

[Top of Page]

8. Applicability of GMP

GMP is expected to apply not only to the final API, but also to intermediate stages of the synthesis. But a significant regulatory burden is incurred by companies when lengthy syntheses are subject to regulatory review and GMP inspection. How far back in the synthesis GMP should reach remains a “gray area”. FDA has accepted that it is unreasonable to apply GMP concepts to the original progenitors of the API, which may have come from plants, mines, oil wells, or sea water. However, “it is reasonable to expect GMP concepts to start to become applicable at that point where a starting material enters a biological or chemical synthesis or series of processing steps where it is known that the end product will be a BPC (Bulk Pharmaceutical Chemical)” [15].

A key issue is therefore the definition of the “starting materials” for each synthesis. According to the Q7A guideline, these are materials which are “incorporated as a significant structural fragment into the structure of the API”. Starting materials may be purchased from an external supplier or contract manufacturer, or manufactured in-house, and accepted on the basis of conformance to an analytical specification; there is no requirement to observe GMP during their manufacture. Thus it is in the API manufacturer's interest to define the starting materials as late in the synthesis as possible. In the past, starting materials were normally expected to be commercially available substances with a variety of uses. Compounds produced specifically for the API synthesis were regarded as intermediates, and therefore subject to GMP. In practice, however, many such compounds were accepted by FDA as starting materials, after review on a case-by-case basis. The official FDA criteria for a starting material were [16]:

> It is a compound whose name, chemical structure, chemical and physical characteristics and properties, and impurity profile are well-defined in the chemical literature, or
> It is obtained by commonly known procedures (this applies principally to starting materials extracted from plants and animals, and to semisynthetic antibiotics).

More recently, commentators have criticized the vagueness of these criteria. PhRMA [17] has suggested as an alternative that a noncommercial substance should be regarded as a starting material if all of the following apply:

> It is a significant structural fragment of the drug substance.
> It has a well-characterized structure and well-defined purity profile (whether or not published in the open literature).
> Specifications exist which ensure appropriate control over downstream processing and drug substance quality.
> It is not the final intermediate, used in the step which generates the API molecule itself.

The argument is that the improvements in analytical technology over the past decade make it more reasonable to expect the quality attributes for starting materials to be readily identified and controlled.

Once the starting materials have been defined, all subsequent synthetic steps should be carried out under GMP, but this need not be at a uniform level:

"Although strict observance of GMPs, approaching or equaling those expected for finished drug products, may be expected in some types of bulk processes, in most others it is neither feasible nor required to apply rigid controls during the early processing steps. In all processes of this type, however, the requirements should be increasingly tightened according to some reasonable rationale. At some logical processing step, usually well before the final finishing operation, appropriate GMPs should be imposed and maintained throughout the remainder of the process." [15].

For example, the production of a very early intermediate may not require the ultralow residue limits on equipment, the process segregation, or the environmental control associated with the final step. While all intermediates should be evaluated and approved before further processing, for the earlier steps this may be performed entirely within the production department, whereas for the later steps the involvement of the quality unit is mandatory. All intermediates should be tested, but for *in situ* intermediates, this may involve no more than an in-process control. It is not necessary to set formal specifications for each intermediate; but companies should identify the critical stages of their syntheses at which specifications would be appropriate. The final intermediate is normally regarded as a major watershed in the synthesis — the point beyond which GMP controls should be significantly increased; here the specifications and tests should be nearly as stringent as those applied to the final API itself.

[Top of Page]

9. GMP in API Process Development

Pharmaceutical companies conduct a substantial amount of research into potential new therapeutic agents. Most candidate new drugs will never be marketed; they are rejected either for lack of safety, lack of efficacy, lack of bioavailability, or insufficient improvement over existing therapies. Nonetheless, they are tested on human volunteers before these decisions can be made. APIs intended for use in such clinical studies are expected to be produced according to appropriate GMP standards.

(Note that where the clinical trial material is produced on a small scale in a laboratory, its manufacture should still be subject to GMP, and not GLP. GLP relates to the conduct of laboratory studies, not to manufacturing on a laboratory scale.)

The application of GMP at the early investigative stage often gives rise to practical difficulties. By their nature, clinical investigations must be carried out before there has been time to develop the optimum manufacturing methods. One of the key features of GMP is process consistency, but this is the hallmark of a relatively mature process, and so is an unrealistic expectation during the development of that process. Often, the early studies use API prepared by the original discovery synthesis, which would not be appropriate for later large-scale manufacturing. It is difficult to set specifications for the starting materials or intermediates in the absence of long processing experience. Also, without validated analytical methods, it is difficult to be sure what impurities are present in the API. In some cases it is difficult to reproduce the physical form of the product on a consistent basis. Clinical APIs will normally be produced in chemical pilot plants, where the reactors are used for numerous processes, often only on one or two occasions each. Thus it is difficult to develop the validated cleaning methods which guarantee against cross-contamination.

There is therefore a need for much more process flexibility at the development stage than could be tolerated in commercial production. There must be more reliance on case-wise scientific judgment than on immutable specifications. But this increased flexibility is mitigated by the more detailed attention that can be given to the production of individual batches, the use of more highly qualified and experienced personnel, the more extensive testing carried out, and the relatively small number of patients involved in the studies.

The Q7A guideline recognizes that not all the controls it recommends are appropriate for the manufacture of clinical APIs, and makes some broad recommendations in Section XIX. This theme has also been developed in greater detail by the API Committee of CEFIC, the European Chemical Industries Council [18]. Their approach to GMP for API Development is based on the following principles:

- The product should be fit for its intended use, and the safety of any subject in the clinical trails should be guaranteed.
- All information relating to the quality of the product or process should be retrievable and reconstructable.
- Quality must be built into the process — the earlier the better.
- During the development of an API the GMP controls should increase.

During early development there may not even be a firm specification for the API itself. The principle generally observed is that its quality should be at least as good — and preferably better — than that used in the preclinical toxicology studies or previous clinical studies. In particular it should contain no new impurities which were not present in those previous batches.

GMP lays particular emphasis on the importance of good record keeping, and if anything this is even more vital in the early project stages (when processes are changing rapidly) than during commercial operations (when the processes are largely fixed). Information generated from the beginning of the project should therefore be constantly documented in development reports. These allow the project to develop in a logical scientific manner, and facilitate the validation of the process when it moves into commercial manufacturing.

But while it is necessary to exercise flexibility with process control issues during API development, there is no reason to be less stringent in areas such as housekeeping, cleanliness, record keeping, equipment maintenance, calibration or staff training. In these areas the standards in the pilot plant should be similar to those observed in a commercial API plant.

[Top of Page]

References

1. P. Uys: Good Manufacturing Practice. *A Guide to Practical Quality Management*, Sigma Press, Pretoria 1994.
2. K. McCormack: *Quality*, Butterworth-Heinnemann Pharmaceutical Engineering Series 2002.
3. I. R. Berry, D. Harpaz (eds): *Validation of Active Pharmaceutical Ingredients*, 2nd ed., IHS Health Group, Englewood, CO 2001.
4. B. K. Immel: "A Brief History of the GMPs for Pharmaceuticals", *Pharm. Technol.* **25** (2001) no. 7, 44 – 52.
5. *Current Good Manufacturing Practice for Finished Pharmaceuticals*, US Code of Federal Regulations, 21CFR Part 211, 1997
6. Food & Drug Administration: Guidance for Industry: *Q7A Good Manufacturing Practice Guidance for Active Pharmaceutical Ingredients*, Rockville, MD 2001.
7. UK Medicines Control Agency: *Rules and Guidance for Pharmaceutical Manufacturers and Distributors*, The Stationery Office, London 1997.
8. Pharmaceutical Inspections Convention: *Regulations on the Manufacture of Pharmaceutical Active Ingredients (2/87)*, 1987.
9. European Chemical Industry Council & European Federation of Pharmaceutical Industries' Associations: *Good Manufacturing Practices for Active Ingredient Manufacturers*, Brussels 1996.
10. Quality Control Bulk Pharmaceuticals Workgroup, PhRMA, *Pharm. Technol*., Dec. 1995, Jan. 1996.
11. ISO 9000: 2000: *Quality Management Systems – Requirements*, ISO, Geneva 2000.
12. http://www.fda.gov/cder
13. Food & Drug Administration: *Guide to Inspections Validation of Cleaning Processes*, Rockville, MD 1993.
14. Food & Drug Administration: *Guideline on Principles of Process Validation*, Rockville, MD 1987.
15. Food & Drug Administration: *Guide to Inspection of Bulk Pharmaceutical Chemicals*, Rockville, MD 1991.
16. Food & Drug Administration: *Guideline for Submitting Supporting Documentation in Drug Applications for the Manufacture of Drug Substances*, Rockville, MD 1987.
17. T. Cupps, B. Fritschel, W. Mavroudakis, M. Mitchell, D. Ridge, J. Wyvratt: *Pharm. Technol.* **27** (2003) no. 2, 34 – 50.
18. APIC, European Chemical Industries Council: *Good Manufacturing Practices in API-Development*, Brussels 1999.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Nonimmobilized Biocatalysts in Industrial Fine Chemical Synthesis

Standard Article

Andreas Liese[1] and Stephan Lütz[2]

[1]Universität Münster, Institut für Biochemie, Münster, Germany

[2]Forschungszentrum Jülich GmbH, Institut für Biotechnologie, Jülich, Germany

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.h17_h01

Article Online Posting Date: January 15, 2004

Recommend to Your Librarian

Save title to My Profile

Email this page

Print this page

BROWSE THIS TITLE

Article Titles A–Z

Topics

SEARCH THIS TITLE

Go

Advanced Product Search

Search All Content

Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Examples of Reactions
2.1. Reductions Catalyzed by Oxidoreductases (E.C. 1)
2.1.1. Ketone Reduction with Whole Cells of *Neurospora crassa* (E.C. 1.1.1.1)
2.1.2. Ketoester Reduction with Cell Extract of *Acinetobacter Calcoaceticus* (E.C. 1.1.1.1)
2.1.3. Enantioselective Reduction with Whole Cells of *Candida sorbophila* (EC 1.1.X.X)
2.1.4. Production of Chiral α-Hydroxy Acids with D-Lactate Dehydrogenase (E.C. 1.1.1.28) from *Leuconostoc mesenteroides*
2.1.5. Converting Ketoacid Acetals to Acetal Amino Acids with L-Phenylalanine Dehydrogenase (PheDH, E.C. 1.4.1.20) from *Thermoactinomyces intermedius*
2.2. Oxidations Catalyzed by Oxidoreductases (E.C. 1)
2.2.1. Alcohol Oxidation with Whole Cells of *Gluconobacter suboxydans* (E.C. 1.1.99.21)
2.2.2. Kinetic Resolution by Oxidation of Primary Alcohols Catalyzed by Whole Cells from *Rhodococcus erythropolis* (E.C. 1.X.X.X)
2.2.3. Hydroxylation of Nicotinic Acid (Niacin) Catalyzed by Whole Cells of *Achromobacter xylosoxidans* (E.C. 1.5.1.13)
2.2.4. Reduction of Hydrogen Peroxide Concentration by Catalase (E.C. 1.11.1.6)
2.2.5. Bioconversion of *N*-Butylglucamine with Whole Cells of *Gluconobacter oxydans* (E.C. 1.1.99.21)
2.3. Hydrolytic Cleavage and Formation of C—O and C—N Bonds by Hydrolases (E.C. 3)
2.3.1. Kinetic Resolution of a Diester by Protease Subtilisin Carlsberg from *Bacillus sp*. (E.C. 3.4.21.62) ,
2.3.2. Kinetic Resolution of α-Amino Acid Amides Catalyzed by Aminopeptidase from *Pseudomonas putida* (E.C. 3.4.1.11)
2.3.3. Production of L-Methionine by Kinetic Resolution with Aminoacylase of *Aspergillus oryzae* (E.C. 3.5.1.14)
2.3.4. Synthesis of β-Lactam Antibiotics Catalyzed by Penicillin Acylase (E.C. 3.5.1.11)
2.3.5. Enantioselective Synthesis of an Aspartame Precursor with Thermolysin from *Bacillus proteolicus* (E.C. 3.4.24.27)
2.3.6. Hydrolysis of Heterocyclic Nitriles by Nitrilase/Hydroxylase from *Agrobacterium sp*. (E.C. 3.5.5.1)
2.3.7. Resolution of 3,3,3-Trifluoro-2-hydroxy-2-methylpropionamide with Amidase from *Klebsiella oxytoca* (E.C. 3.5.X.X)
2.4. Formation of C—C, C—O, and C—N Bonds by Lyases (E.C. 4)
2.4.1. Hydrocyanation of *m*-Phenoxybenzaldehyde with Oxynitrilase from *Hevea brasiliensis* (E.C. 4.1.2.39)
2.4.2. Synthesis of Carnitine Catalyzed by Carnitine Dehydratase in Whole Cells (E.C. 4.2.1.89)
2.4.3. Synthesis of L-Dopa Catalyzed by Tyrosine Phenol Lyase from *Erwinia herbicola* (E.C. 4.1.99.2)

2.4.4. Synthesis of Nicotinamide Catalyzed by Nitrile Hydratase from *Rhodococcus erythropolis* (EC 4.2.1.84) ,
2.4.5. Synthesis of Mandelic Acid with Nitrilase (E.C. 4.2.1.84) ,
2.5. Epimerization of Glucosamine Catalysed by Epimerase from *E. coli* (E.C. 5.1.3.8)
3. Outlook

This article focuses on the use of nonimmobilized biocatalysts for synthesis of fine chemicals in industry. More than 20 processes using isolated enzymes or suspensions are presented.

[Top of Page]

1. Introduction

Enzymes as homogeneous catalysts have found widespread application for the synthesis of fine chemicals, mostly due to their selectivity [1], [2]. Although some enzymatic approaches are clearly superior to classical chemical methods, only a few processes have reached larger scale industrial application [3, 7]. While the use of immobilized biocatalysts ⟶ is well known, this article focuses on the use of nonimmobilized biocatalysts in industrial fine-chemical synthesis. The term “nonimmobilized“ refers to biocatalysts not bound or absorbed on a support, i.e., isolated enzymes and suspensions of whole cells.

There are many prejudices against biotransformations in the chemical industry, the two most important of which are:

1. Biocatalysts are too expensive
2. Biocatalysts only work under mild conditions

The first is only true in part. If the cost per mole or unit weight is calculated then they are certainly expensive. For example, penicillin amidase costs $ 10 000/kg on a bulk scale. However, the cost contribution of penicillin amidase in the splitting of penicillin G is only $ 1 per kilogram of product [8]. In L-aspartic acid production the cost contribution of aspartase is even lower, namely, $ 0.1/kg. This demonstrates that not the absolute catalyst cost but the cost contribution of the catalyst to the final product cost has to be considered and compared. This is also true for chemical catalysts, e.g., the bulk price of BINAP is $ 40 000/kg [8]. Important parameters influencing the cost contribution are the total turnover number (moles of product per mole of catalyst) and the turnover frequency (moles of product per mole of catalyst and unit time).

The second prejudice, that biocatalysts only work in aqueous phase with low concentrations of starting material, is also only true in parts. The natural environment is in general the aqueous phase and ambient temperature. But the examples described below demonstrate that biocatalysts can also be applied in emulsions or even pure organic solvents (Table 1). Here very high concentrations can be reached, e.g., in the case of acrylamide up to 5.6 M.

Table 1. Highest concentrations applied in industrial biotransformations

E.C. no.	Enzyme	Substrate	Concentration	Medium
1.1.99.21	D-sorbitol dehydrogenase	1-amino-D-sorbitol (N-protected)	1.00 M	aqueous
4.2.1.2	fumarase	fumaric acid	1.00 M	aqueous
4.3.1.1	L-aspartase	fumaric acid	1.00 M	aqueous
3.4.21.62	subtilisin	phenylalanine isopropyl ester	1.20 M	aqueous
3.1.1.3	lipase	1-phenylethylamine	1.65 M	MTBE
3.5.2.6	β-lactamase	γ-lactam	1.83 M	aqueous
4.2.1.84	nitrile hydratase	adiponitrile	2.01 M	aqueous/organic
4.3.1.1	L-aspartase	fumaric acid	2.00 M	aqueous
4.1.1.12	aspartate β-decarboxylase	aspartic acid	2.50 M	aqueous
3.1.1.25	lactonase	pantolactone	2.69 M	aqueous
3.1.1.3	lipase	palmitic acid	3.10 M	2-propanol
3.1.1.3	lipase	cyclopentenyl ester	4.16 M	aqueous/organic
4.2.1.84	nitrile hydratase	acrylonitrile	5.60 M (product)	aqueous
4.3.1.5	L-phenylalanine ammonia lyase	*trans*-cinnamic acid	9.31 M (NH_3)	aqueous

What is generally considered as a drawback can often prove to be an advantage of biocatalysts over chemical catalysts, since the environmentally benign reaction conditions may lower the process costs by reducing expensive waste due to a higher atom efficiency and lower energy costs [9], [10]. Also, in most cases the number of reaction steps can be reduced

by avoiding protection/deprotection sequences.

[Top of Page]

2. Examples of Reactions

The examples presented here are mostly taken from [3]. Newer processes have been added as far as information was available. Information on the synthesis of the substrates and application of the products can be found in the references on the specific process. If available, the whole synthetic route in which the biotransformation is integrated is given. The enzymes involved, either in a whole cell or as an isolated catalyst, are mainly from the groups of oxidoreductases (E.C. class 1) and hydrolases (E.C. class 3). There are a few examples for lyases (E.C. class 4) and one example for an isomerase (E.C. class 5). The majority of processes involving oxidoreductases use whole cells due to the problem of cofactor regeneration. The examples are sorted in the order of the main enzyme classes (EC). The letter **E** denotes the biotransformation step in the synthetic schemes.

2.1. Reductions Catalyzed by Oxidoreductases (E.C. 1)

2.1.1. *Ketone Reduction with Whole Cells of Neurospora crassa (E.C. 1.1.1.1) [11-14]*

The key step in the synthesis of Trusopt, a topically active treatment for glaucoma, is the enantioselective reduction of 5,6-dihydro-6-methyl-4*H*-thieno[2,3*b*]thiopyran-4-one-7,7-dioxide (Fig. 1).

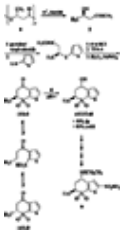

Figure 1. Synthesis of Trusopt (**5**) via enzymatic reduction (Astra-Zeneca)

1: biologically derived homopolymer, **2**: (*R*)-methyl 3-hydroxybutyrate, **3**: 5,6-dihydro-6-methyl-4*H*-thieno[2,3*b*]thiopyran-4-one 7,7-dioxide, **4**: 5,6-dihydro-4-hydroxy-6-methyl-4*H*-thieno[2,3*b*]thiopyran-4-one-7,7-dioxide, **5**: Trusopt (MK-0507), **E**: alcohol dehydrogenase (whole cells of *Neurospora crassa*).

TFAA = trifluoroacetic acid.

The biological route overcomes the problem of incomplete inversion of the *cis* alcohol in the chemical synthesis (Fig. 2). The reaction is carried out below pH 5 to prevent epimerization of the (6*S*)-methylketosulfone in aqueous media.

Figure 2. Chemical synthesis of Trusopt

DMS = dimethyl sulfide, TFAA = trifluoroacetic acid.

(*R*)-3-Hydroxymethyl butyrate (**1**), which is responsible for the stereochemistry of the methyl group in the sulfone ring, can be produced by depolymerization of biopolymers, e.g., Biopol from Zeneca, which is a natural polyester produced by some microorganisms as a storage compound.

2.1.2. *Ketoester Reduction with Cell Extract of Acinetobacter calcoaceticus (E.C. 1.1.1.1) [15-17]*

The biotransformation (Fig. 3) is an alternative to the chemical synthesis via the chlorohydrin and selective hydrolysis of the acyloxy group (Fig. 4). After final fractional distillation this synthesis has an overall yield of 41 %. The biotransformation has a yield of 92 %. The diketoester can be obtained as shown in Figure 5.

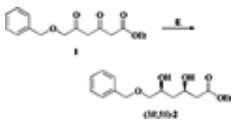

Figure 3. Synthesis of a key intermediate for anticholesterol drugs

1: 6-benzyloxy-(3*R*,5*S*)-dioxohexanoic acid ethyl ester, **2**: 6-benzyloxy-(3*R*,5*S*)-dihydroxyhexanoic acid ethyl ester, **E**: alcohol dehydrogenase from *Acinetobacter calcoaceticus*.

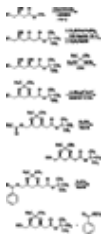

Figure 4. Chemical synthesis of a key intermediate for anticholesterol drugs

CSA = camphorsulfonic acid, LiHMDS = lithium hexamethyldisilazide, NMP = *N*-methyl-2-pyrrolidone.

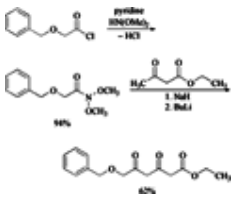

Figure 5. Synthesis of 6-benzyloxy-3,5-dihydroxyhexanoic acid ethyl ester

6-Benzyloxy-(3*R*,5*S*)-dihydroxyhexanoic acid ethyl ester is a key chiral intermediate for anticholesterol drugs that act by inhibition of hydroxymethylglutaryl coenzyme A (HMG CoA) reductase.

2.1.3. *Enantioselective Reduction with Whole Cells of Candida sorbophila (EC 1.1.X.X) [18-20]*

Here the biotransformation (Fig. 6) is preferred over the chemical reduction with commercially available asymmetric catalysts (BH_3- or noble-metal-based), since with the chemical catalysts the desired high enantiomeric excess (e.e. > 98 %, 99.8 % after purification) is not achievable. Since the ketone has only very low solubility in the aqueous phase, 1 kg ketone is added as solution in 4 L of 0.9 M H_2SO_4 to the bioreactor. The bioreduction is essentially carried out in a two-phase system consisting of the aqueous phase and small droplets composed of substrate and product. The downstream processing consists of multiple extraction steps with methyl ethyl ketone and precipitation induced by pH titration of the pyridine functional group (pK_a = 4.66) with NaOH. The (*R*)-amino alcohol is an important intermediate for the synthesis of β-3-agonists that can be used for obesity therapy and to decrease the level of associated type II diabetes, coronary artery disease, and hypertension.

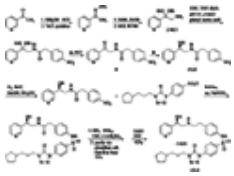

Figure 6. Synthesis of a key intermediate for a β-3-agonist (Merck Research Laboratories)

1: 2-(4-nitrophenyl)-*N*-[2-oxo-2-(3-pyridyl)ethyl]acetamide, **2**: (*R*)-*N*-[2-hydroxy-2-(3-pyridyl)ethyl]-2-(4-nitrophenyl) acetamide, **3**: β-3-agonist, **F**: Dehydrogenase (whole cells of *Candida sorbophila*).

EDC = *N*-(3-dimethylaminopropyl)-*N*-ethylcarbodiimide hydrochloride MTBE = methyl *tert*-butyl ether.

2.1.4. *Production of Chiral α-Hydroxy Acids with D-Lactate Dehydrogenase (E.C. 1.1.1.28) from Leuconostoc mesenteroides [21]*

In a membrane reactor (*R*)-3-(4-fluorophenyl)-2-hydroxypropionic acid is produced in a continuous process with high enantioselectivity (e.e. > 99.9 %) and a high space–time yield (560 g L^{-1} d^{-1}). The biotransformation was favored since an alternative chemical route yielded the alcohol in varying selectivity (78–92 % e.e.).

D-Lactate dehydrogenase (D-LDH) from *Leuconostoc mesenteroides* was favored over D-LDH from *Staphylococcus epidermidis* for economic reasons. The cofactor NADH was regenerated by oxidizing formate to carbon dioxide with formate dehydrogenase from *Candida boidinii* (Fig. 7). The product was obtained in good yield (80–90 %) and converted to the methyl ester for the determination of optical purity (e.e. > 99.9 %). The process is designed for multikilogram scale. The product is a precursor for a potential rhinovirus protease inhibitor.

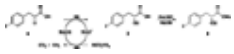

Figure 7. Synthesis of a chiral α-hydroxy acid (Pfizer Global Research and Development)

1: 4-Fluorophenyl-2-hydroxypropionic acid, **2**: (*R*)-3-(4-fluorophenyl)-2-hydroxypropionic acid, **3**: (*R*)-3-(4-fluorophenyl)-2-hydroxypropionic acid methyl ester, **E1**: D-lactate dehydrogenase (E.C. 1.1.1.28) from *Leuconostoc mesenteroides*, **E2**: formate dehydrogenase from *Candida boidinii*.

2.1.5. *Converting Ketoacid Acetals to Acetal Amino Acids with L-Phenylalanine Dehydrogenase (PheDH, E.C. 1.4.1.20) from Thermoactinomyces intermedius [22-24]*

The amino acid acetal is a precursor for Omapatrilat, an antihypertensive drug, which inhibits the angiotensin-converting enzyme (ACE) and neutral endopeptidase (NEP). The synthesis is carried out in 16 000 L batch reactions with heat-dried *E. coli* cells containing the cloned and overexpressed PheDH from *Thermoactinomyces intermedius* and the formate dehydrogenase from *C. boidinii* for cofactor regeneration (Fig. 8). Yield (97 %), conversion (97 %), and optical purity of the product (e.e. > 99 %) are very high.

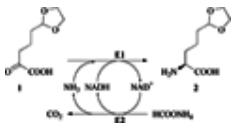

Figure 8. Synthesis of an acetal amino acid (Bristol Myers Squibb)

1: 5-(1,3-dioxolan-3-yl)-2-oxopentanoic acid (keto acid acetal), **2**: (*S*)-allysine ethylene acetal (acetal amino acid), **E1**: phenylalanine dehydrogenase (PheDH) from *Thermoactinomyces intermedius*, **E2**: formate dehydrogenase from *C. boidinii*.

2.2. Oxidations Catalyzed by Oxidoreductases (E.C. 1)

2.2.1. *Alcohol Oxidation with Whole Cells of Gluconobacter suboxydans (E.C. 1.1.99.21) [25-27]*

In 1923 the bacterium *Acinetobacter suboxydans* was isolated, and starting in the mid-1930s was used for the industrial oxidation of L-sorbitol to L-sorbose in the Reichstein–Grüssner synthesis of vitamin C [28]. Bayer uses the same type of reaction, but instead of *Acinetobacter* the bacterium *Gluconobacter suboxydans* is used in the oxidation of *N*-protected 6-

amino-L-sorbitol to the corresponding 6-amino-L-sorbose, which is an intermediate in the production of miglitol (Fig. 9). 1-Desoxynojirimycin is produced by chemical intramolecular reductive amination of 6-amino-L-sorbose. In contrast, the published chemical synthesis of 1-desoxynojirimycin and its derivatives requires multiple steps and laborious protecting-group chemistry. Miglitol and derivatives thereof are pharmaceuticals for treatment of carbohydrate metabolism disorders (e.g., diabetes mellitus).

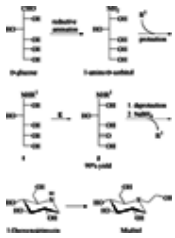

Figure 9. Synthesis of key intermediate for miglitol (Bayer)

1: 1-amino-D-sorbitol, **2**: 6-amino-L-sorbose (*N*-protected), **E**: D-sorbitol dehydrogenase (whole cells of *Gluconobacter suboxydans*).

2.2.2. *Kinetic Resolution by Oxidation of Primary Alcohols Catalyzed by Whole Cells from Rhodococcus erythropolis (E.C. 1.X.X.X) [29, 20, 31]*

(*R*)-Isopropylideneglycerol is a useful C_3 synthon in the synthesis of (*S*)- β-blockers, e.g., (*S*)-metoprolol. (*R*)-Isopropylideneglyceric acid can also be used as starting material for the synthesis of biologically active products. The resolution is carried out by selective microbial oxidation of the *S* enantiomer (Fig. 10). The chemical synthesis of (*R*)-isopropylideneglycerol starts from nonnatural L-mannitol or from L-ascorbic acid (Fig. 11). In comparison to the biotransformation, here a stoichiometric quantity of lead tetraacetate is needed.

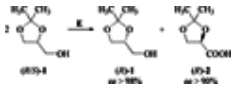

Figure 10. Synthesis of (*R*)-isopropylideneglycerol and (*R*)-isopropylideneglyceric acid (International Biosynthetics)

1: isopropylideneglycerol, **2**: (*R*)-isopropylideneglyceric acid, **E**: oxidase (whole cells from *Rhodococcus erythropolis*).

Figure 11. Chemical synthesis of (*R*)-isopropylideneglycerol from nonnatural L-mannitol or from L-ascorbic acid

2.2.3. *Hydroxylation of Nicotinic Acid (Niacin) Catalyzed by Whole Cells of Achromobacter xylosoxidans (E.C. 1.5.1.13) [32-34]*

6-Hydroxynicotinate is a versatile building block, chiefly in the synthesis of modern insecticides. The 6-hydroxynicotinate-producing strain (Fig. 12) was found by accident when precipitated white crystals of 6-hydroxynicotinate were found in the mother liquor of a niacin-producing chemical plant. The second enzyme of the nicotinic acid pathway, decarboxylating 6-hydroxynicotinate hydroxylase, is strongly inhibited at niacin concentrations above 1 %, whereas niacin hydroxylase is unaffected. In contrast to the biotransformation the chemical synthesis of 6-substituted nicotinic acids is difficult and expensive due to the separation of byproducts that are produced by nonregioselective hydroxylations.

Figure 12. Synthesis of 6-hydroxynicotinate (Lonza)

1: niacin = nicotinic acid = pyridine-3-carboxylate, **2**: 6-hydroxynicotinate = 6-hydroxypyridine-3-carboxylate, **E**: nicotinic acid hydroxylase (whole cells of *Achromobacter xylosoxidans*).

2.2.4. *Reduction of Hydrogen Peroxide Concentration by Catalase (E.C. 1.11.1.6) [35]*

During oxidative coupling of 1-nitrotoluene to give dinitrodibenzyl (DNDB) hydrogen peroxide is formed as a byproduct. It is not possible to decompose H_2O_2 by adding heavy-metal catalysts because only incomplete conversion is reached, and subsequent process steps with DNDB are problematic due to contamination with heavy-metal catalyst. The biotransformation is only used to decompose the undesired side product H_2O_2 (Fig. 13). The enzyme of choice is catalase derived from a microbial source, which has advantages over bovine catalase since the activity remains constant over a broad pH range from 6.0 to 9.0, temperatures up to 50 °C are tolerated, and salt concentrations up to 25 % do not affect the enzyme stability. The reaction is carried out in a cascade of three continuous stirred-tank reactors (Fig. 14). The H_2O_2 concentration is reduced from 7000 to < 200 ppm in the product solution. The third vessel is aerated with nitrogen to degas the product solution. The dinitrodibenzyl is used as a pharmaceutical intermediate.

Figure 13. Degradation of hydrogen peroxide (Novartis).

1: 1-nitrotoluene, 2: dinitrodibenzyl (DNDB), E: catalase from a microbial source.

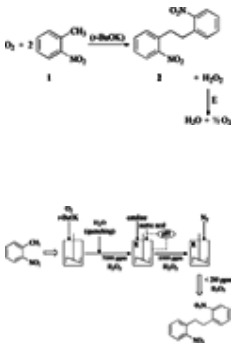

Figure 14. Flow scheme of dinitrodibenyl synthesis

2.2.5. *Bioconversion of N-Butylglucamine with Whole Cells of Gluconobacter oxydans (E.C. 1.1.99.21) [36]*

N-Butyldeoxynojirimycin is an inhibitor of glycosidase activity. A precursor for this potential therapeutic agent against retroviral infections, 6-deoxy-6-butylaminosorbose, is synthesized by oxidation with whole cells of *Gluconobacter oxydans* (Fig. 15). The enzyme involved is sorbitol dehydrogenase (SDH, see Section Alcohol Oxidation with Whole Cells of *Gluconobacter suboxydans* (E.C. 1.1.99.21)). The key to this process is the temperature of the biotransformation, due to degradation of the sorbose product. The selectivity could be enhanced by lowering the process temperature to 12–15 °C, and the product was obtained in 95 % yield. With optimized conditions the process was scaled up to a 5500 L reactor.

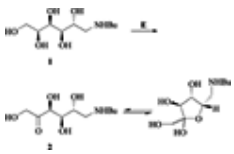

Figure 15. Biosynthesis of 6-deoxy-6-butylaminosorbose (Pharmacia)

1: *N*-butylglucamine, **2**: 6-deoxy-6-butylaminosorbose, **E**: whole cells of *Gluconobacter oxydans*.

2.3. Hydrolytic Cleavage and Formation of C—O and C—N Bonds by Hydrolases (E.C. 3)

2.3.1. *Kinetic Resolution of a Diester by Protease Subtilisin Carlsberg from Bacillus sp. (E.C. 3.4.21.62) [37], [38]*

(*R*)-(2-Methylpropyl)butanedioic acid 4-ethyl ester is used as chiral building block for potential collagenase inhibitors (e.g., Ro 31-9790) for the treatment of osteoarthritis. A 20 % emulsion of the diester in 30 mM aqueous $NaHCO_3$ is treated with Protease L660 or Alcalase 2.5L (9 % with respect to the racemic diester; Fig. 16). The unconverted *S* diester can be extracted into a solvent such as toluene and racemized by heating the anhydrous extract with catalytic amounts of sodium ethanolate. The resulting racemic diester can be recycled, and thus the overall yield is improved from 45 to 87 %. The reaction was repeatedly carried out on a 200 kg scale with respect to the racemic diester. The enzyme is highly stereoselective even at high substrate concentrations (20 %). The chemoenzymatic route starting from the cheap bulk agents maleic anhydride and isobutylene replaced the existing chemical synthesis for bulk amounts (Fig. 17).

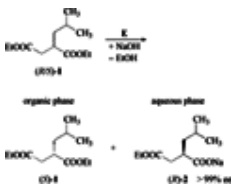

Figure 16. Synthesis of (*R*)-(2-methylpropyl)butanedioic acid 4-ethyl ester (Hoffmann La-Roche)

1: (*R*)-(2-Methylpropyl)butanedioic acid diethyl ester, **2**: (*R*)-(2-methylpropyl)butanedioic acid 4-ethyl ester, Na form, **E**: hydrolase (subtilisin Carlsberg from *Bacillus sp.*).

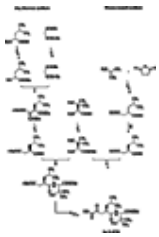

Figure 17. Comparison of drug discovery and process research route to Ro 31-9790

2.3.2. *Kinetic Resolution of α-Amino Acid Amides Catalyzed by Aminopeptidase from Pseudomonas putida (E.C. 3.4.1.11) [39-42]*

Enantiomerically pure α-amino acids are intermediates in the synthesis of antibiotics and are used for parenteral nutrition and as food and feed additives (⟶ Amino Acids). Examples are D-phenylglycine and 4-hydroxyphenylalanine for semisynthetic β-lactam antibiotics and L-phenylalanine for the peptidic sweetener aspartame. DSM also used this process to produce L-homophenylalanine, a potential precursor for several ACE inhibitors.

The α-amino amides as substrates for this enantiospecific, biocatalytic amide hydrolysis can be readily obtained from the appropriate aldehydes by Strecker synthesis (Fig. 18). As whole-cell catalyst *Pseudomonas putida* is applied, which accepts a wide range of substrates. Subsequent to the biotransformation benzaldehyde is added to precipitate the D-

amide Schiff base, which can be easily isolated by filtration. An acidification step leads to the D-amino acid. The L-amino acid can be reused after racemization, so that a theoretical yield of 100 % D-amino acid is possible. The same process can be used for the synthesis of 100 % L-amino acids by racemizing the Schiff base of the D-amide with small amounts of base in organic solvents.

Figure 18. Production of L- and D- α-amino acids by kinetic resolution of α-amino acid amides (DSM)

1: aldehyde, **2**: aminonitrile, **3**: α-amino acid amide, **4**: α-amino acid methyl ester, **5**: α-amino acid, **6**: α-amino acid amide, **7**: Schiff base of α-amino acid amide, **E**: aminopeptidase from *Pseudomonas putida*.

By using in vivo protein engineering not only mutant strains of *Pseudomonas putida* exhibiting L-amidase and D-amidase, but also amino acid amide racemase activities were obtained. Using these mutants a convenient synthesis of α-amino acids with 100 % yield would be possible with a one-cell system. It is noteworthy that only α-H substrates can be used. By screening, a new biocatalyst of the strain *Mycobacterium neoaurum* was found which is capable of converting α-substituted amino acid amides.

2.3.3. *Production of L-Methionine by Kinetic Resolution with Aminoacylase of Aspergillus oryzae (E.C. 3.5.1.14) [43-46]*

The *N*-acetyl-D,L-amino acid precursors are conveniently accessible through acetylation of D,L-amino acids with acetyl chloride or acetic anhydride in a Schotten–Baumann reaction or by amidocarbonylation [47]. For the acylase reaction Co^{2+} is added as metal effector to yield an increased operational stability of the enzyme (Fig. 19). The unconverted acetyl-D-methionine is racemized by acetic anhydride in alkali, and the racemic *N*-acetyl-D,L-methionine is reused. The racemization can also be carried out in a molten bath or by an acetyl amino acid racemase. The L-methionine product is recovered by crystallization, because L-methionine is much less soluble than the acetyl substrate. The production is carried out in a continuous stirred-tank reactor. A polyamide ultrafiltration membrane with a cutoff of 10 000 dalton retains the enzyme and thus decouples the residence times of catalyst and reactants. L-Methionine is produced with an e.e. > 99.5 % and a yield of 80 % with a capacity of > 300 t/a. Degussa produces several proteinogenic and nonproteinogenic amino acids in the same way, e.g., L-alanine, L-phenylalanine, α-aminobutyric acid, L-valine, L-norvaline, and L-homophenylalanine.

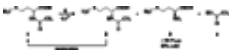

Figure 19. Biocatalytic production of L-methionine by kinetic resolution (Degussa)

1: *N*-acetyl-D,L-methionine, **2**: *N*-acetyl-D-methionine, **3**: L-methionine, **4**: acetic acid, **E**: aminoacylase from *Aspergillus oryzae*.

2.3.4. *Synthesis of β-Lactam Antibiotics Catalyzed by Penicillin Acylase (E.C. 3.5.1.11) [48-51]*

The penicillin acylases do not accept charged amino groups. Therefore phenylglycine itself cannot be used at a pH value at which the carboxyl function is necessarily uncharged, because the amino group will then be charged. To reach nonequilibrium concentrations of the product, the substrate must be activated as an ester or amide (Fig. 20). Thus the amino group can be partly uncharged at the optimal pH value of the enzyme. In biological systems ATP delivers the activation energy. Using the same synthetic pathway, alternatively to 7-ADCA and 7-ACCA also 6-APA derivatives can be synthesized.

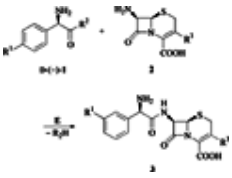

Figure 20. Synthesis of β-lactam antibiotics (Chemferm)

1a: phenylglycine amide (PGA; R^1 = H, R^2 = NH_2), **1b**: phenylglycine methyl ester (PGM; R^1 = H, R^2 = OMe), **1c**: hydroxyphenylglycine amide (HPGA; R^1 = OH, R^2 = NH_2), **1d**: hydroxyphenylglycine methyl ester (HPGM; R^1 = OH, R^2 = OMe), **2a**: 7-aminodeacetoxy-3-methylcephalosporanic acid (7-ADCA; R^3 = Me), **2b**: 7-aminodeacetoxy-3-chlorocephalosporanic acid (7-ACCA; R^3 = Cl), **3a**: Cefaclor (R^1 = H, R^3 = Cl), **3b**: Cephalexin (R^1 = H, R^3 = Me), **3c**: Cefadroxil (R^1 = OH, R^3 = Me).

The established chemical synthesis started from benzaldehyde and included fermentation of penicillin (Fig. 21). The process consists of ten steps and produces 30–40 kg of waste per kilogram product. The waste contained methylene chloride, other solvents, silylating agents, and numerous products from side-chain protection and acylating promoters. In comparison the chemoenzymatic route needs only six steps, of which three are biocatalytic. The biotransformation **E1** in Figure 21 is carried out with penicillin amidase from *Escherichia coli* (E.C. 3.5.1.11), and **E2** with aminopeptidase from *Pseudomonas putida* (E.C. 3.4.1.11).

Figure 21. Comparison of chemical and biocatalytic synthesis of cefalaxin

E1: penicillin amidase, **E2**: amino peptidase, **E3**: penicillin acylase.

7-ADCA = 7-aminodeacetoxy-3-methylcephalosporanic acid (see Fig. 20).

2.3.5. *Enantioselective Synthesis of an Aspartame Precursor with Thermolysin from Bacillus proteolicus (E.C. 3.4.24.27) [52], [53]*

Since the reaction (Fig. 22) is limited by equilibrium the products must be removed from the reaction mixture to reach high yields. Therefore, an excess of racemic phenylalanine methyl ester, which is inert to the reaction, is added. The carboxylic anion of the protected aspartame forms a poorly soluble adduct D-Phe-OCH_3, which precipitates from the reaction mixture and can be removed easily by filtration. The final steps of the process are the separation of D-Phe ester, removal of protecting groups, and racemization of the formed L-amino acid. α-Aspartame is produced with > 99.9 % e.e. World capacity is about 10 000 t/a, of which about 2500 t/a is produced by enzymatic coupling. Since the bacterial strain was found in the Rokko Hot Spring in central Japan it is thermally stable (up to 60 °C).

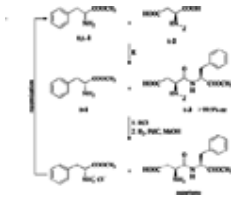

Figure 22. Biocatalytic synthesis of aspartame (Holland Sweetener Company)

1: phenylalanine methyl ester, **2**: aspartic acid (protected), **3**: α-aspartame (protected), thermolysin from *Bacillus proteolicus*.

The main problem in chemical synthesis by coupling of Z-Asp anhydride with L-Phe-OCH_3 is the formation of β-aspartame as byproduct. This isomer has a bitter taste and must be completely removed from the α isomer. The advantages of the enzymatic route are:

1. no β isomer is produced,
2. the enzyme is completely stereoselective, so that racemic mixtures of the substrate or the appropiate enantiomer of the amino acid can be used,
3. no racemization occurs during synthesis, and
4. the reaction takes place in aqueous media under mild conditions.

2.3.6. *Hydrolysis of Heterocyclic Nitriles by Nitrilase/Hydroxylase from Agrobacterium sp. (E.C. 3.5.5.1) [54-56]*

5-Hydroxypyrazine-2-carboxylic acid (Fig. 23) is a versatile building block in the synthesis of new antituberculosis agents, e.g., 5-chloropyrazine-2-carboxylic acid esters. The regioselective hydroxylation of pyrazine-2-carboxylic acid is catalyzed by a hydroxylase (**E2**, E.C. 1.5.1.13), which is contained in suspended whole cells of *Agrobacterium sp*. The biomass is separated by ultrafiltration (cutoff 10 kDa) after biotransformation. 5-Hydroxypyrazine-2-carboxylic acid is precipitated from the permeate by acidification with sulfuric acid to pH 2.5.

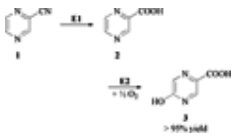

Figure 23. Biocatalytic synthesis of 5-hydroxypyrazine-2-carboxylic acid (Lonza)

1: 2-cyanopyrazine, **2**: pyrazine-2-carboxylic acid, **3**: 5-hydroxypyrazine-2-carboxylic acid, **E1/E2**: nitrilase/hydroxylase (whole cells of *Agrobacterium sp*., which contains both enzymes.

In contrast to the biotransformation the chemical synthesis of 5-substituted pyrazine-2-carboxylic acid leads to a mixture of 5- and 6-substituted pyrazinecarboxylic acids and requires multiple steps.

2.3.7. *Resolution of 3,3,3-Trifluoro-2-hydroxy-2-methylpropionamide with Amidase from Klebsiella oxytoca (E.C. 3.5.X.X) [57]*

A novel heat-stable amidase from *Klebsiella oxytoca* has been isolated, characterized, and cloned. For safety reasons, the amidase is overexpressed in *E. coli*. As homogeneous biocatalyst heat-treated whole cells of *E. coli* are used for the hydrolysis the methylpropionamide to give the *R* acid and the *S* amide. The amide can be hydrolyzed chemically to the *S* acid, which could be obtained as product as well (Fig. 24). Optical purity is greater than 98 % for the acid, which is produced on a 100 kg scale. The products are intermediates for the synthesis of a number of potential pharmaceuticals, which include ATP-sensitive potassium channel openers and diabetes remedies.

Figure 24. Resolution of 3,3,3-trifluoro-2-hydroxy-2-methylpropionamide (Lonza)

1: 3,3,3-trifluoro-2-hydroxy-2-methylpropionamide, **2**: 3,3,3-trifluoro-2-hydroxy-2-methylpropionic acid, **E**: amidase from

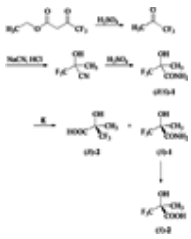

Klebsiella oxytoca.

2.4. Formation of C—C, C—O, and C—N Bonds by Lyases (E.C. 4)

2.4.1. *Hydrocyanation of m-Phenoxybenzaldehyde with Oxynitrilase from Hevea brasiliensis (E.C. 4.1.2.39) [58]*

Whole cells of an undisclosed host microorganism overproducing the oxynitrilase from *Hevea brasiliensis* are used in a biphasic solvent system of aqueous buffer and methyl *tert*-butyl ether for the enantioselective hydrocyanation of *m*-phenoxybenzaldehyde (Fig. 25). This biphasic system has been proven beneficial and applicable on an industrial scale. The product is an important precursor for pyrethroids.

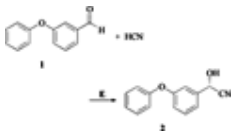

Figure 25. Hydrocyanation of *m*-phenoxybenzaldehyde (DSM)

1: *m*-phenoxybenzaldehyde, **2**: (*S*)-*m*-phenoxybenzaldehyde cyanohydrin, **E**: (*S*)-oxynitrilase from *Hevea brasiliensis*.

2.4.2. *Synthesis of Carnitine Catalyzed by Carnitine Dehydratase in Whole Cells (E.C. 4.2.1.89) [59-61]*

L-Carnitine is used in infant, health, sport, and geriatric nutrition. The biotransformation is catalyzed by carnitine dehydratase in whole cells (Fig. 26). (*R*)-Carnitine is produced with > 99.5 % conversion of 4-butyrobetaine and > 99.5 % e.e.. The mutant strain has blocked L-carnitine dehydrogenase and excretes the accumulated product. The purified enzyme could not be used for the biotransformation due to its high instability. Apart from usual batch fermentations, continuous production is also feasible since the cells go into a “maintenance state” with high metabolic activity and low growth rate. The cells can be recycled after separation from the fermentation broth by filtration. A chemical resolution process with L-tartaric acid that was developed at Lonza was no longer competitive to the biotechnological route. A more attractive route would be the Ru/BINAP-catalyzed asymmetric hydrogenation of 4-chloroacetoacetate (Fig. 26), in which an e.e. of 97 % is achieved.

Figure 26. Comparison of chemical and biocatalytic syntheses of carnitine (Lonza)

1: 4-butyrobetaine, **2**: carnitine, **E**: carnitine dehydratase in whole cells of *E coli*.

2.4.3. *Synthesis of L-Dopa Catalyzed by Tyrosine Phenol Lyase from Erwinia herbicola (E.C. 4.1.99.2) [62-65]*

L-Dopa is used for the treatment of Parkinsonism that is caused by a lack of L-dopamine and its receptors in the brain. L-Dopamine is synthesized in organisms by decarboxylation of L-3,4-dihydroxyphenylalanine (L-dopa). Since L-dopamine can not pass the blood–brain barrier L-dopa is applied in combination with dopadecarboxylase inhibitors to avoid formation of L-dopamine outside the brain. Ajinomoto produces L-dopa by this lyase biotransformation (Fig. 27) with suspended whole cells in a fed batch reactor on a scale of 250 t/a. Much earlier Monsanto has successfully scaled up the chemical synthesis of L-dopa (Fig. 28).

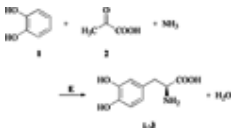

Figure 27. Biocatalytic synthesis of L-dopa (Ajinomoto)

1: catechol, **2**: pyruvic acid, **3**: L-dopa, **E**: tyrosine phenol lyase from *Erwinia herbicola*.

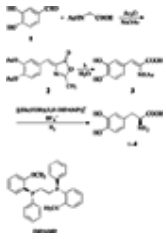

Figure 28. Chemical synthesis of L-dopa (Monsanto)

1: vanillin, **2**: azlactone, **3**: (*Z*)-enamide, **3**: L-dopa.

The enantioselective hydrogenation of 3,4-dihydroxy-*N*-acetylamino cinnamic acid is catalyzed by a cationic Rh complex of the biphosphine DIPAMP, in which the enantioselectivity is introduced by the chiral phosphine [66], [67]. The hydrogenation proceeds quantitatively with 94 % e.e.. The optically pure L-dopa is separated from the catalyst by crystallization.

2.4.4. *Synthesis of Nicotinamide Catalyzed by Nitrile Hydratase from Rhodococcus erythropolis (E.C. 4.2.1.84) [34], [68]*

Nicotinamide (vitamin B3) is used as vitamin supplement for food and animal feed. The same strain is used as in the industrial production of acrylamide [3]. The biotransformation is carried out on a scale of 3000 t/a (Fig. 29).

Figure 29. Comparison of chemical and biocatalytic syntheses of nicotinamide (Lonza)

E: nitrile hydratase (whole cells of *Rhodococcus erythropolis*).

In contrast to the chemical alkaline hydrolysis of 3-cyanopyridine with 4 % of nicotinic acid byproduct (96 % yield) the biotransformation gives complete selectivity and no acid or base is required. The continuous biotransformation is operated at low temperature and atmospheric pressure. In contrast to the old synthesis route for nicotinamide at Lonza the new pathway is environmentally friendly and safe. Only one organic solvent is used throughout the whole process in four highly selective, continuous, and catalytic reactions. The process water, NH_3, and H_2 are recycled.

2.4.5. *Synthesis of Mandelic Acid with Nitrilase (E.C. 4.2.1.84) [5], [69]*

By kinetic resolution of racemic mandelonitrile (Fig. 30) mandelic acid is produced on a multiton per year scale. By choosing the appropriate reaction conditions (i.e., pH) the remaining enantiomer in the resolution is re-racemized in the presence of cyanide ions due to an equilibrium between hydroxy nitrile and aldehyde.

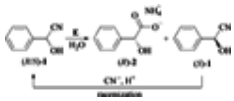

Figure 30. Resolution of mandelonitrile (BASF)

1: mandelonitrile, **2**: mandelic acid, **E**: nitrilase.

2.5. Epimerization of Glucosamine Catalysed by Epimerase from *E. coli* (E.C. 5.1.3.8) [70, 71]

N-Acetyl-D-mannosamine serves as *in situ* generated substrate for the synthesis of *N*-acetylneuraminic acid. Since *N*-acetyl-D-mannosamine is quite expensive it is synthesized from *N*-acetyl-D-glucosamine by epimerization at C-2. This biotransformation is integrated in the production of *N*-acetylneuraminic acid (Neu5Ac).

By using *N*-acylglucosamine 2-epimerase it is possible to start with the inexpensive *N*-acetyl-D-glucosamine instead of *N*-acetyl-D-mannosamine (Fig. 31). The epimerase is used for the *in situ* synthesis of *N*-acetyl-D-mannosamine (ManNAc). Since the equilibrium is on the side of the starting material, the reaction is driven by the subsequent biotransformation of ManNAc together with pyruvate to give Neu5Ac.

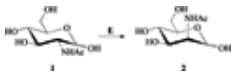

Figure 31. Biocalytic epimerization of glucosamine to mannosamine (Marukin Shoyu)

1: *N*-acetyl-D-glucosamine, **2**: *N*-acetyl-D-mannosamine, **E**: *N*-acylglucosamine 2-epimerase from *E. coli*.

The *N*-acylglucosamine 2-epimerase is cloned from porcine kidney, transformed and overexpressed in *Escherichia coli*. To reach maximal activity ATP and Mg^{2+} must be added. Since the whole synthesis is reversible high GlcNAc concentrations are used.

The chemical epimerization of GlcNAc is carried out by Glaxo. The equilibrium of the chemical epimerization is on the side of *N*-acetyl-D-glucosamine (GlcNAc:ManNAc = 4:1). After neutralization and addition of 2-propanol GlcNAc precipitates. In the remaining solution a ratio of GlcNAc:ManNAc = 1:1 is reached. After evaporation to dryness and and extraction with methanol the GlcNAc:ManNAc ratio is shifted to 1:4.

[Top of Page]

3. Outlook

Biotransformations have gained more and more importance in industry. Classical chemical steps in the industrial synthesis have been successfully replaced. For industrial application biocatalysts need not be immobilized. Moreover the potential of biocatalysis is still increasing. By improved screening methods new catalysts will be detected and made available in large amounts by cloning and overexpression. Directed evolution will be used to improve properties such as stability or selectivity [73], [74]. Metabolic engineering will be used to analyze and remove bottlenecks in the metabolism or to create novel biocatalysts [75].

[Top of Page]

References

1. C. H. Wong, “Enzymes in Organic-Synthesis — Oxidoreductions”, *J. Chem. Soc. Perkin Trans*. 1995, 967 – 978.
2. K. M. Koeller, C. H. Wong, “Enzymes for Chemical Synthesis”, *Nature* **409** (2001) 232 – 240. Links
3. A. Liese, K. Seelbach, C. Wandrey, *Industrial Biotransformations*, Wiley-VCH, Weinheim, Germany, 2000.
4. A. Liese: “Industrial Applications and Processes using Biocatalysis” in K. Drauz, H. Waldmann, S. M. Roberts (eds.) *Enzyme Catalysis in Organic Synthesis*, 2nd ed., VCH-Wiley, Weinheim, Germany, 2002, 1419 – 1460.
5. A. Schmid, J. S. Dordick, B. Hauer, A. Kiener, M. Wubbolts, B. Witholt, “Industrial Biocatalysis Today and Tomorrow”, *Nature* **409** (2001) 258 – 268. Links
6. A. J. J. Straathof, S. Panke, A. Schmid, “The Production of Fine Chemicals by Biotransformations, *Curr. Opin. Biotechnol*. **13** (2002) 548 – 556. Links
7. R. N. Patel, “Microbial/Enzymatic Synthesis of Chiral Intermediates for Pharmaceuticals”, *Enzyme Microbiol. Technol*. **31** (2002) 804 – 826.
8. J. D. Rozzell, “Commercial Scale Biocatalysis: Myths and Realities”, *Bioorg. Med. Chem*. **7** (1999) 2253 – 2261. Links
9. J. H. Clark, “Green Chemistry”, *Green Chem*. **1** (1999) 1 – 8.
10. R. A. Sheldon, “Consider the Environmental Quotient”, *Chemtech* **24** (1994) 38 – 47.
11. A. J. Blacker, R. A. Holt, “Development of a Multistage Chemical and Biological Process to an Optically Active Intermediate for an Anti-Glaucoma Drug” in A. N. Collins, G. N. Sheldrake, J. Crosby (eds.): *Chirality In Industry II*, Wiley, New York 1997, pp. 246 – 261.
12. T. J. Blacklock, P. Sohar, J. W. Butcher, T. Lamanec, E. J. J. Grabowski, “An Enantioselective Synthesis of the Topically-Active Carbonic Anhydrase Inhibitor MK-0507: 5,6-dihydro-(*S*)-4-(ethylamino)-(*S*)-6-methyl-4*H*-thieno[2,3-*b*] thiopyran-2-sulfonamide 7,7-dioxide hydrochloride”, *J. Org. Chem*. **58** (1993) 1672 – 1679. Links
13. R. A. Holt, “Microbial Asymmetric Reduction in the Synthesis of a Drug Intermediate”, *Chim. Oggi* **9** (1996) 17 – 20.
14. Zeneca, US 5580764, 1996 (R. A. Holt, S. R. Rigby).
15. E. R. Squibb &Sons, EP 0569998A2, 1993 (R. N. Patel, C. G. McNamee).
16. R. N. Patel, A. Banerjee, C. G. McNamee, D. Brzozowski, R. L. Hanson, L. J. Szarka, “Enantioselective Microbial Reduction of 3,5-Dioxo-6-(benzyloxy) Hexanoic Acid, Ethyl Ester, *Enzyme Microbiol. Technol*. **15** (1993) 1014 – 1021.
17. S. Y. Sit, R. A. Parker, I. Motoc, W. Han, N. Balasubramanian, “Synthesis, Biological Profile, and Quantitative Structure-Activity Relationship of a Series of Novel 3-Hydroxy-3-methylglutaryl Coenzyme A Reductase Inhibitors”, *J. Med. Chem*. **33** (1990) no. 11, 2982 – 2999. Links
18. Merck, US 5846791, 1998 (M. Chartrain, J. Chung, C. N. Roberge).
19. J. Chung, G. Ho, M. Chartrain, C. Roberge, D. Zhao, J. Leazer, R. Farr, M. Robbins, K. Emerson, D. Mathre, J. McNamara, D. Hughes, E. Grabowski, P. Reider, “Practical Chemoenzymatic Synthesis of a Pyridylethanolamino Beta-3 Adrenergic Receptor Antagonist–Stereospecific Beta-3 Agonist Used in Obesity Therapy, Produced by Yeast-Mediated Asymmetric Reduction of a Ketone”, *Tetrahedron Lett*. **40** (1999) 6739 – 6743. Links
20. M. Chartrain, C. Roberge, J. Chung, J. McNamara, D. Zhao, R. Olewinski, R. Hunt, P. Salmon, D. Roush, S. Yamazaki, T. Wang, E. Grabowski, B. Buckland, R. Greasham, “Asymmetric Bioreduction of (2-(4-nitro-phenyl)-*N*-(2-oxo-2-pyridin-3-yl-ethyl)-acetamide) to its Corresponding (*R*) Alcohol [(*R*)-*N*-(2-hydroxy-2-pyridin-3-yl-ethyl)-2-(4-nitro-phenyl)-acetamide] by Using *Candida sorbophila* MY 1833”, *Enzyme Microbiol. Technol*. **25** (1999) 489 – 496.
21. J. H. Tao, K. McGee, “Development of a Continuous Enzymatic Process for the Preparation of (*R*)-3-(4-Fluorophenyl)-2-hydroxypropionic Acid”, *Org. Process Res. Dev*. **6** (2002) 520 – 524. Links
22. R. N. Patel, “Biocatalytic Synthesis of Intermediates for the Synthesis of Chiral Drug Substances”, *Curr. Opin. Biotechnol*. **12** (2001) 587 – 604. Links
23. R. N. Patel, “Enzymatic Synthesis of Chiral Intermediates for Drug Development”, *Adv. Synth. Catal*. **343** (2001) 527 – 546. Links
24. R. N. Patel, “Enzymatic Synthesis of Chiral Intermediates for Omapatrilat, an Antihypertensive Drug”, *Biomol. Eng*. **17** (2001) 167 – 182. Links
25. Bayer AG, DE 2834122 A1, 1978 (G. Kinast, M. Schedel).
26. Bayer AG, DE 2853573 A1, 1978 (G. Kinast, M. Schedel).
27. G. Kinast, M. Schedel, “Vierstufige 1-Desoxynojirimycin-Synthese mit einer Biotransformation als zentralem Reaktionsschritt”, *Angew. Chem*. **93** (1981) 799 – 800.
28. T. Reichstein, H. Grüssner, “Eine ergiebige Synthese der L-Ascorbinsäure (Vitamin C)” *Helv. Chim. Acta* **17** (1934) 311 – 328. Links
29. J. Crosby, “Synthesis of Optically Active Compounds: A Large Scale Perspective”, *Tetrahedron* **47** (1991) 4789 – 4846. Links
30. Gist Brocades NV and Shell Int. Research, EU 244912, 1987 (M. A. Bertola, H. S. Koger, G. T. Phillips, A. F. Marx, V. P. Claassen).
31. G. Hirth, W. Walther, “Synthesis of the (*R*)- and (*S*)-Glycerol Acetonides. Determination of the Optical Purity”, *Helv. Chim. Acta* **68** (1985) 1863. Links
32. H. G. Kulla, “Enzymatic Hydroxylations in Industrial Application”, *Chimia* **45** (1991) 81 – 85. Links
33. Lonza, EP 0152948 A2, 1995 (P. Lehky, H. Kulla, S. Mischler).
34. M. Petersen, A. Kiener, “Biocatalysis–Preparation and Functionalization of N-Heterocycles”, *Green Chem*. **2** (1999) 99 – 106.

35. U. Onken, E. Schmidt, T. Weissenrieder, “Enzymatic H_2O_2 Decomposition in a Three-Phase Suspension”, Ciba-Geigy, *International Conferenence on Biotechnology for Industrial Production of Fine Chemicals*, 93rd event of the EFB, Zermatt, Switzerland, 29.09.1996.
36. B. H. Landis, J. K. McLaughlin, R. Heeren, R. W. Grabner, P. T. Wang, “Bioconversion of N-Butylglucamine to 6-Deoxy-6-butylamino Sorbose by *Gluconobacter oxydans*”, *Org. Process Res. Dev*. **6** (2002) 547 – 552. Links
37. B. Wirz, T. Weisbrod, H. Estermann, “Enzymatic Reactions in Process Research — The Importance of Parameter Optimization and Workup”, *Chimica Oggi*. **14** (1995) 37 – 41.
38. S. Doswald, H. Estermann, E. Kupfer, H. Stadler, W. Walther, T. Weisbrod, B. Wirz, W. Wostl, “Large Scale Preparation of Chiral Building Blocks for The P3 Site of Renin Inhibitors”, *Bioorg. Med. Chem*. **2** (1994) 403 – 410. Links
39. J. Kamphuis, E. M. Meijer, W. H. J. Boesten, T. Sonke, W. J. J. van den Tweel, H. E. Schoemaker, “New Developments in the Synthesis of Natural and Unnatural Amino Acids” in D. S. Clark, D. A. Estell (eds.): *Enzyme Engineering XI*, Ann. N.Y. Acad. Sci. **672** (1992) 510 – 527.
40. H. E. Schoemaker, W. H. J. Boesten, B. Kaptein, H. F. M. Hermes, T. Sonke, Q. B. Broxterman, W. J. J. van den Tweel, J. Kamphuis, “Chemo-enzymatic Synthesis of Amino Acids and Derivatives”, *Pure Appl. Chem*. **64** (1992) 1171 – 1175. Links
41. W. J. J. van den Tweel, T. J. G. M. van Dooren, P. H. de Jonge, B. Kaptein, A. L. L. Duchateau, J. Kamphuis, “*Ochrobacterium anthropi* NCIMB 40321: A New Biocatalyst with Broad-Spectrum L-Specific Amidase Activity”, *Appl. Microbiol. Biotechnol*. **39** (1993) 296 – 300.
42. J. Kamphuis, H. F. M. Hermes, J. A. M. van Balken, H. E. Schoemaker, W. H. J. Boesten E. M. Meijer, “Chemo-enzymatic of Enantiomerically Pure Alpha-H and Alpha-Alkyl Alpha-Amino Acids and Derivates” in G. Lubec, G. A. Rosenthal (Eds.): *Amino Acids: Chemistry, Biology, Medicine*, ESCOM Science Puplishers, Leiden, 1990, pp. 119 – 125.
43. A. S. Bommarius, K. Drauz, H. Klenk, C. Wandrey, “Operational Stability of Enzymes–Acylase-Catalyzed Resolution of *N*-Acetyl Amino Acids to Enantiomerically Pure L-Amino Acids”, *Ann. N. Y. Acad. Sci*. **672** (1992) 126 – 136. Links
44. W. Leuchtenberger, M. Karrenbauer, U. Plöcker, “Scale-up of an Enzyme Membrane Reactor Process for the Manufacture of L-Enantiomeric Compounds” (Enzyme Engineering 7), *Ann. N. Y. Acad. Sci*. **434** (1984) 78. Links
45. C. Wandrey, E. Flaschel, “Process Development and Economic Aspects in Enzyme Engineering. Acylase L-Methionine System” in T. K. Ghose, A. Fiechter, N. Blakebrough (Eds.): *Advances in Biochemical Engineering* **12**, Springer-Verlag, Berlin, 1979, pp. 147 – 218.
46. Degussa, US 4304858, 1981 (C. Wandrey, R. Wichmann, W. Leuchtenberger, M. R. Kula).
47. M. Beller, M. Eckert, W. Moradi, “First Amidocarbonylation with Nitriles for the Synthesis of *N*-Acyl Amino Acids”, *Synlett*, 1999, 108. Links
48. A. Bruggink, “Biocatalysis and Process Integration in the Synthesis of Semi-Synthetic Antibiotics”, *Chimia* **50** (1996) 431 – 432. Links
49. A. Bruggink, E. C. Roos, E. de Vroom, “Penicillin Acylase in the Industrial Production of β-Lactam Antibiotics”, *Org. Proc. Res. Dev*. **2** (1998) 128 – 133.
50. Gist-Brocardes, US 5470717, 1995 (K. Clausen).
51. O. Hernandez-Justiz, R. Fernandez-Lafuente, J. M. Terrini, Guisan, “Use of Aqueous Two-Phase Systems for In Situ Extraction of Water Soluble Antibiotics during their Synthesis by Enzymes Immobilized on Porous Supports”, *Biotech. Bioeng*. **59** (1998) no. 1, 73 – 79.
52. K. Oyama, “The Industrial Production of Aspartame” in A. N. Collins, G. N. Sheldrake, J. Crosby (Eds.): *Chirality in Industry*, Wiley, New York, 1992, 237 – 247.
53. Holland Sweetener Company, EP 0768384, 1996 (T. Harada, Y. Shinnanyo-shi, S. Irino, Y. Kunisawa, K. Oyama).
54. Lonza, EP 578137 A1, 1994 (A. Kiener).
55. A. Kiener, J. -P. Roduit, A. Tschech, A. Tinschert, K. Heinzmann, “Regiospecific Enzymatic Hydroxylations of Pyrazinecarboxylic Acid and a Practical Synthesis of 5-Chloropyrazine-2-Carboxylic Acid”, *Synlett* **10** (1994) 814 – 816.
56. M. Wieser, K. Heinzmann, A. Kiener, “Bioconversion of 2-Cyanopyrazine to 5-Hydroxypyrazine-2-carbolic Acid with *Agrobacterium sp*. DSM 6336”, *Appl. Microbiol. Biotechnol*. **48** (1997) 174 – 180. Links
57. N. M. Shaw, A. Naughton, K. Robins, A. Tinschert, E. Schmid, M. L. Hischier, V. Venetz, J. Werlen, T. Zimmermann, W. Brieden, P. de Riedmatten, J. P. Roduit, B. Zimmermann, R. Neumuller, “Selection, Purification, Characterisation, and Cloning of a Novel Heat-Stable Stereo-specific Amidase from Klebsiella oxytoca, and its Application in the Synthesis of Enantiomerically Pure (R)- and (S)-3,3,3-Trifluoro-2-hydroxy-2-methylpropionic Acids and (S)-3,3,3-Trifluoro-2-hydroxy-2-methylpropionamide”, *Org. Process Res. Dev*. **6** (2002) 497 – 504. Links
58. H. Gröger, “Enzymatic Routes to Enantiomerically Pure Aromatic α-Hydroxy Carboxylic Acids: A Further example for the Diversity of Biocatalysis”, *Adv. Synth. Catal*. **343** (2001) 547 – 558. Links
59. M. Kitamura, T. Ohkuma, H. Takaya, R. Noyori, “A Practical Asymmetric Synthesis of Carnitine”, *Tetrahedron Lett*. **29** (1988) 1555 – 1556. Links
60. J. Macy, H. Kulla, G. Gottschalk, “H_2 -Dependent Anaerobic Growth of *Escherichia coli* on L-Malate: Succinate Formation”, *J. Bacteriol*. **125** (1976) 423 – 428. Links
61. Th. P. Zimmermann, K. T. Robins, J. Werlen, F. W. Hoeks, “Bio-transformation in the Production of L-Carnitine” in A. N. Collins, G. N. Sheldrake, J. Crosby (eds.): *Chirality in Industry*, Wiley, New York, 1997, 287 – 305.
62. Ajinimoto, JP 5123177A, 1993 (T. Tsuchida, Y. Nishimoto, T. Kotani, K. Iiizumi).
63. D. J. Ager, *Handbook of Chiral Chemicals*, Marcel Dekker, New York, 1999.
64. Ajinomoto, JP 1010995A, 1989 (A. Yamamoto, K. Yokozeki, K. Kubota).

65. H. Yamada, “Screening of Novel Enzymes for the Production of Useful Compounds”, in K. Kieslich, C. P. van der Beek, J. A. M. de Bont, W. J. J. van den Tweel (eds.): *New Frontiers in Screening for Microbial Biocatalysis, Studies in Organic Chemistry 53*, Elsevier, Amsterdam, 1998, pp. 13 – 17.
66. B. Cornils, W. A. Herrmann, R. Schlögl, C.-H. Wong, *Catalysis from A to Z*, Wiley-VCH, Weinheim, Germany, 2000.
67. W. Knowles, M. Sabacky, B. Vineyard, D. Weinkauff, “Asymmetric Hydrogenation with a Complex of Rhodium and a Chiral Bisphosphine”, *J. Am. Chem. Soc*. **97** (1975) 2567 – 2568. Links
68. J. Heveling, “Catalysis at Lonza: From Metallic Glasses to Fine Chemicals”, *Chimia* **50** (1996) 114 – 118. Links
69. BASF, DE 19848129, 2000 (M. Ress-Löschke, T. Friedrich, B. Hauer, R. Mattes, D. Engels).
70. U. Kragl, D. Gygax, O. Ghisalba, C. Wandrey, “Enzymatic Process for Preparing *N*-Acetylneuraminic Acid”, *Angew. Chem. Int. Ed. Engl*. **30** (1991) 827 – 828.
71. I. Maru, J. Ohnishi, Y. Ohta, Y. Tsukada, “Simple and Large-Scale Production of *N*-Acetylneuraminic Acid from *N*-Acetyl-D-glucosamine and Pyruvate Using *N*-acyl-D-Glucosamine 2-epimerase and *N*-Acetylneuraminate lyase”, *Carbohydr. Res*. **306** (1998) 575 – 578. Links
72. I. Maru, Y. Ohta, K. Murata, Y. Tsukada, “Molecular Cloning and Indentification of N-Acyl-D-glucosamine 2-epimerase from Porcine Kidney as a Renin-Binding Protein”, *Biol. Chem*. **271** (1996) 16294 – 16299.
73. F. H. Arnold, “Combinatorial and Computational Challenges for Biocatalyst Design”, *Nature* **409** (2001) 253 – 257. Links
74. U. Bornscheuer, M. Pohl, “Improved Biocatalysts by Directed Evolution and Rational Protein Design”, *Curr. Opin. Chem. Biol*. **5** (2000) 137 – 142 Links
75. M. Chartrain, P. Salmon, D. Robinson, B. Buckland, “Metabolic Engineering and Directed Evolution for the Production of Pharmaceuticals”, *Curr. Opin. Biotechnol*. **11** (2000) 209 – 214. Links

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Alkali Metal Cyanides

Andreas Rubo, CyPlus GmbH, Hanau-Wolfgang, Federal Republic of Germany
Raf Kellens, DSM, Geleen, The Netherlands
Jay Reddy, Sasol, Johannesburg, South Africa
Norbert Steier, CyPlus GmbH, Wesseling, Federal Republic of Germany
Wolfgang Hasenpusch, Degussa AG, Frankfurt, Federal Republic of Germany

1. Properties 1
2. Production, Packaging, and Transport 2
2.1. Classical Cyanide Production Processes 3
2.2. Cyanide Production Processes 3
2.3. Energy Consumption 4
2.4. Packaging and Transport 4
2.5. Responsible Care, Product Stewardship, Sustainable Development, Certification 5
3. Emissions and Consumptions 5
4. Quality Specifications, Impurities . . . 7
5. Uses 7
5.1. Gold Mining 7
5.2. Chemical Synthesis 8
5.3. Electroplating of Metals 8
5.4. Hardening of Metals 9
6. Economic Aspects 9
7. Best Available Techniques (BAT) . . . 10
7.1. Processes 10
7.2. Procedures 10
7.3. Toxicology and Occupational Health . 15
8. References 15

1. Properties

Pure sodium cyanide NaCN [*143-33-9*] and potassium cyanide KCN [*151-50-8*] are white crystalline solids. They are readily soluble in water and can be dissolved to prepare aqueous solutions with ca. 30 % (NaCN) and ca. 40 % (KCN) concentration at 20 °C. These cyanides are often called "alkaline cyanides" because they render water alkaline pH when dissolved. This is due to a dissolution equilibrium in which free alkali is formed from these salts of a weak acid and a strong base:

$$MCN + H_2O \rightleftarrows HCN + MOH$$

Apart from dissolution in water, HCN can be formed and released from alkali metal cyanide solutions and solids by contact with a large range of weak to strong acids. The liberation of HCN is pH-dependent (Fig. 1).

On contact with acids the white color of solid cyanides can change, and colorless solutions can turn light yellow to dark brown. The intensity of the color change

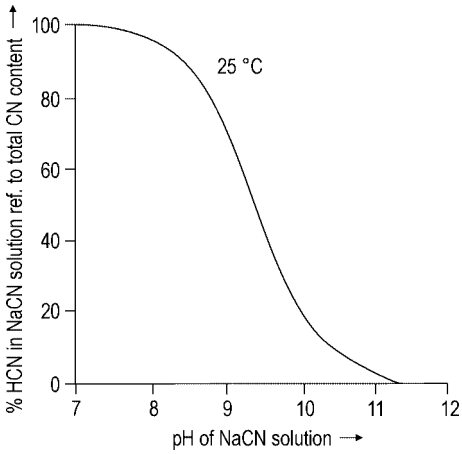

Figure 1. HCN/cyanide equilibrium versus pH

depends on a variety of parameters. This behavior is due to polymerization of HCN under alkaline conditions. As this behavior is of importance for the quality of cyanide solutions, these products must be stabilized by addition of excess alkali. The alkali (normally NaOH or KOH) acts as a pH buffer whenever cyanide comes into contact with smaller amounts of acids and neutral-

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.i01_i01

izes them. In this sense alkali metal cyanides are sensitive to any contact with humid air which contains carbon dioxide, which is a weak acid in aqueous solution. Only when additional stabilization with alkali has been carried out is the acid neutralized to form carbonate and the cyanide protected from decomposition by the acid.

Further physical properties of NaCN and KCN are listed in Table 1.

Table 1. Physical properties of NaCN and KCN

Physical quantity	NaCN	KCN
M_r, relative molecular mass	49.015	65.119
mp, melting point	561.7 °C (98 wt %)	634.5 °C (98 wt %)
bp, boiling point	1500 ± 10 °C	
ϱ, density	1.620 g/cm^3 (6 °C, rhombic) 1.595 g/cm^3 (20 °C, cubic) 1.19 g/cm^3 (850 °C, fused)	1.553 g/cm^3 (20 °C, cubic) 1.56 g/cm^3 (25 °C, cubic)
c_p, specific heat capacity	1.667 kJ kg^{-1} K^{-1} (273.1 K) 17.950 kJ kg^{-1} K^{-1} (288.0 K) 1.402 kJ kg^{-1} K^{-1} (298.6 K)	1.00 kJ kg^{-1} K^{-1} (25 – 72 °C)
H, enthalpy (25 °C)	−89.9 kJ/mol	−112.63 kJ/mol
ΔH_f, heat of fusion	314 kJ/kg	225 kJ/kg
ΔH_v, heat of vaporization	3185 kJ/kg	
p_v, vapor pressure	0.10 kPa (800 °C) 1.65 kPa (1000 °C) 11.98 kPa (1200 °C) 39.10 kPa (1350 °C)	
Solubility in 100 g of		
Ethanol, 100 %	1.235 g (25 °C)	0.57 g (19.5 °C)
Ethanol, 95 %	2.445 g (25 °C)	
Methanol	6.44 g (15 °C) 4.58 g (25 °C) 4.10 g (67.4 °C)	4.91 g (19.5 °C)
NH_3, liquid	58 g (−31 °C)	4.55 g (−33.9 °C)

2. Production, Packaging, and Transport

Sodium and potassium cyanide are produced in the same way in Europe and worldwide: gaseous or liquid HCN reacts with a solution of sodium or potassium hydroxide. The formation of cyanides is chemically a simple neutralization reaction between a weak acid (HCN) and a strong base (OH^-). The neutralization reaction has an energy release of 460 kJ/mol, which leads to additional heating of the reaction solution and saves energy.

The efficiency of the neutralization reaction is very high, and the yield based on HCN is close to 100 %. Thus all of the HCN used is transferred into the end products NaCN or KCN, and only minimal losses of HCN occur.

The reaction results in the formation of a dissolved salt. The raw solution can either be concentrated or diluted with water, depending on what kind of intermediate product form is produced. Normally, the aim is to form a concentrated solution with ca. 30 % cyanide. Commercial solid sodium and potassium cyanide can be crystallized by evaporation of excess water.

Worldwide most alkali metal cyanides are produced by utilizing HCN as cyanide source. The production of HCN (→ Cyano Compounds, Inorganic, Chap. 1.2) is, therefore, an important intermediate step for the production of alkali cyanides.

Andrussow process for HCN production:

$$CH_4 + NH_3 + 1.5O_2 \rightarrow HCN + 3H_2O \quad \Delta H = -474\,kJ/mol$$

BMA process for HCN production:

$$CH_4 + NH_3 \rightarrow HCN + 3H_2 \qquad \Delta H = +252\,kJ/mol$$

In the production of acrylonitrile:

$$2\,C_3H_6 + 2\,NH_3 + 3\,O_2 \rightarrow 2\,CH_2CHCN + 6\,H_2O$$

HCN is formed as a byproduct:

$$C_3H_6 + 2\,NH_3 + 8\,O_2 \rightarrow 2\,HCN + 8\,H_2O + 4\,CO_2$$

There are patented cyanide production processes available which directly form solid material in a fluidized-bed reactor using an alkali solution and HCN gas as raw materials. Also processes are patented that use a cyanide solution as the cyanide source, and solid cyanide is produced by a drying process in a fluidized-bed reactor.

2.1. Classical Cyanide Production Processes

The following processes are termed classical as they are not applied to any significant extent in Europe presently but were used in earlier cyanide production.

In the *Castner – Keller process* sodium metal reacts with charcoal and ammonia at around 600 °C in a melt:

$$Na + C + NH_3 \rightarrow NaCN + 3/2\ H_2$$

This process was mainly developed by Degussa and applied there until 1971.

The *calcium cyanamide process* is very similar to the Castner process, but uses calcium carbonate as the carbon source for the cyanide ion.

In the *"blood process"* potassium or sodium carbonate was fused with potassium hexacyanoferrate(II), for which mainly animal blood and wastes were used as raw materials.

2.2. Cyanide Production Processes

Cyanides are made by melt processes or by processes in aqueous solution using HCN as the cyanide source. Most cyanide is produced nowadays using solution processes. Only where (electrical) energy costs are sufficiently low can melt processes be competitive.

Melt processes do not apply any cyanide sources as raw materials, but produce the cyanides from non-cyanide components. The result of the processes is a solid cyanide with different kinds of impurities, e.g., carbon black from using carbon as raw material.

Cyanides can be produced in batch and in continuous operation. Liquid cyanide solutions are easily pumped as the viscosity is moderate even at the highest achievable concentration.

Figure 2 shows a flow sheet for NaCN production.

Production of Cyanide Solution. The production of the cyanide solution (NaCN, KCN) is the basic first production step. Further processing of the resulting raw cyanide solution usually consists of concentration of the solution by evaporation of excess water, cooling, filtering, and storage. The exact ratios of cyanide in water are strongly dependent on the water balance. When a concentrated NaOH solution (50 %) and pure HCN (99 %) are used a cyanide solution with ca. 46 % sodium cyanide is formed which needs to be further diluted with water to give a commercial product.

Figure 2. Example of a solid cyanide production plant

Solutions are often directly used in cases where the site of industrial application is close enough to the cyanide production site to keep the transport of solution competitive (energy for evaporation of the excess water, formulation of the solid material). This is, therefore, usually only applied if the application site can directly use a concentrated cyanide solution (30 wt %). This is, for example, the case for gold mining operations in Nevada, South Africa, and Australia. Cyanide solution is normally shipped in tank trucks to the gold mine, subsequently stored in liquid tanks at the site, and diluted for further use, if necessary.

Many gold mines are too far away from the raw material production site to be supplied with liquid material and, therefore, need solid NaCN.

Cyanide solutions normally contain approx. 24 – 30 % NaCN. Cyanide solutions transported at cold times of the year or in cold regions may contain lower cyanide concentrations to prevent crystallization during transport and storage in unheated devices.

In cases where an application requires HCN as a liquid or a gas and no HCN production is available on site, cyanide solutions or solids can be used as the source. They need only be acidified, and the HCN is distilled and taken to production. Normally, it is not stored intermediately.

Production of Solid Cyanides. Solid cyanide is made from solution by crystallization of solid cyanides from the liquid, separation of the wet solid with further drying and mechanical handling to form stable solid product forms. Figure 3 shows a plant for the production of solid cyanides. The preferred solid forms are bricks, granules, tablets, and powder. Bricks are made by compacting of the wet filter cake with subsequent drying and also by first drying the crystals followed by compacting. Bricks are also made from granules. Powder can be made by grinding and subsequent sieving of fines material.

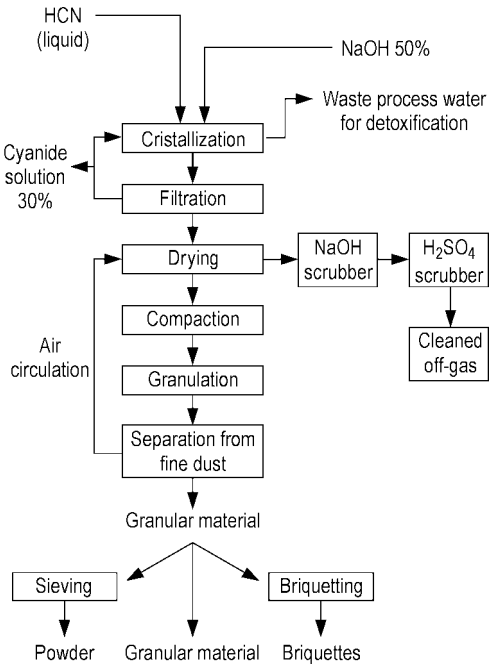

Figure 3. Flow chart of the main production steps for cyanide production (NaCN as example)

The cyanide content in solids is usually very high and specified for normal applications in a range of higher than 98 % and even higher than 99 %.

2.3. Energy Consumption

The energy costs for cyanide production are mainly due to electricity and steam and correspond to approx. 1.3 – 1.5 t of steam per ton of solid 100 % NaCN or KCN.

2.4. Packaging and Transport [5, 6]

Production includes packaging of cyanides for the safe storage and transport. Normally cyanides are filled into metal drums holding between 50 and 100 kg or 1 t big bags which are transported in wooden boxes. All packaging must have UN registration for the packaging of hazardous materials. Returnable packaging for multiple use is an important alternative to the use of one-way packaging. Both one-way packaging and returnable packaging for solid material are used worldwide. Another alternative for transport is the so-called solid to liquid system (SLS). Here, solid cyanides are filled into a unit containing approx. 20 t of material. This unit is transported to the customer's site and the cyanide is dissolved directly at the user's facility and pumped to a storage tank. The transportation unit can be reused and refilled with solid material many times. The preferred type of packaging often depends on the specific transport conditions, which are due to technical, economical, and environmental considerations.

Figure 4 shows storage of solid cyanide in drums.

Transport is regulated by national and international regulations and laws. Road transport of liquid cyanides is allowed in various European countries. However liquid cyanide is preferably shipped by rail cars within Europe. Only certified companies with a special permit for the transportation of dangerous goods are approved. Some European countries require special permits for the transport of cyanides on the road, e.g., Italy. Each transport in Europe is also required to have a transport emergency card with relevant information on cyanide on board.

Figure 4. Storage facility for solid cyanide in drums

In Germany, it is allowed to ship cyanides and acids together in the same shipment (trucks). But for safety reasons cyanides and acids should not be transported and stored together.

Transportation of cyanide solution is mostly favored in cases where transport routes to the end-users are considered to be safe and the distance short enough that transport of liquid products is economic compared to the alternative of solid cyanide. Also the customer must be in the position to directly use the solution.

Production of solid material from cyanide solutions is mostly the best alternative when the product is transported over long distances, when the transport of liquid is not allowed according to national regulations, and when the customers are prepared to store and use solid material.

2.5. Responsible Care, Product Stewardship, Sustainable Development, Certification [7–19]

Production facilities can be certified according to international standards referring to product and management system quality including the relevant aspects of (workers') health, safety, and environmental management or any other systems that lead to similar results. This includes certification according to ISO 9001, 14001, or any other management system that is applicable to manage company and stakeholder needs sufficiently.

Cyanide producers who supply cyanide to gold mines are increasingly being asked to fulfill the requirements of the cyanide producers verification audit protocol according to the International Cyanide Management Code (ICMC) provisions. The overall basis for this kind of certification is the International Cyanide Management Code for the gold mining industry.

All cyanide producers supply material safety data sheets (MSDS) to their customers covering all specific information needed for the safe handling and use of their products. Information and training programmes are in place for traders, end-users, and transport companies.

Transport route analysis from the production facility to the sites of the end-users (mainly for the mining industry) are carried out by producers.

Major European cyanide producers exchange information on safety, health, and environmental aspects of their products an a regular basis within the CEFIC Cyanide Sector Group. This group has, for example, developed a Mutual Aid Scheme (MAS) for transportation incidents with cyanides and published the "Guidelines for Storage, Handling, and Distribution of Alkali Cyanides".

3. Emissions and Consumptions [20–26]

Hazardous Emissions from cyanide production can be emissions of HCN and NH_3 gases and of cyanides, the last-named mainly in wastewater or as particles in the workplace.

NH_3 and HCN can be detected by humans due to their specific odor at low detection limits. These limits are far below any workplace limits. Therefore, people may be warned even without any technical monitoring devices. However, a significant number of persons are not able to detect HCN, and a much smaller number can not detect NH_3 by its smell. Hence, in accordance with standard workplace regulations, monitoring devices must be installed, mainly for the detection of HCN and also for NH_3.

The production of cyanide solutions is operated under closed conditions where no HCN can leak from the units. As a result the overall emission of HCN from cyanide production plants in Europe and elsewhere are below the European and national regulation limits (e.g.,TA Luft in Germany) and below 10 kg HCN per year for a 20 000 t/a production facility.

NH_3 emissions may result from heating cyanides and cyanide solutions. NH_3 formation from cyanide production can result from the hydrolysis of cyanides at higher temperature, which can result in the formation of cyanate. The formation and the subsequent emission of NH_3 to the environment is very low and below reportable limits, i.e., less than 15 kg NH_3 per year.

Detoxification of Cyanides. The detoxification of cyanides can be carried out by using different kinds of oxidizing agents, such as hypochlorite or peroxides. Pure oxygen is not suitable. Using hypochlorite creates a significant salt load of the effluent and can create additional contributions to AOX load. Using H_2O_2 based technology does not create any of these byproducts and can be seen as the best choice for cyanide detoxification at a cyanide production plant.

$$NaCN + OCl^- \rightarrow OCN^- + NaCl$$

$$CN^- + H_2O_2 \rightarrow OCN^- + H_2O$$

Cyanide which is dissolved in process wastewater can be treated by the above-mentioned methods, but preferably peroxides are used (Fig. 5).

Wastewater. The process solution can directly be fed to a biological treatment plant under predefined conditions. Also diluted cyanide effluent streams may directly be added to a biological wastewater treatment, where bacteria break down the cyanide to nontoxic substances and use the chemical as feed.

As a result of detoxification the free cyanide in the treated water is far below 0.01 ppm. The total amount of cyanide transferred to the environment from a 20 000 t/a NaCN plant normally is below 50 kg/a. These small cyanide emissions which enter, e.g., rivers at a very low concentration do not pose harm to fish and the environment and are much likely also detoxified further by microorganisms.

Excess water from concentrating cyanide solutions is evaporated and either recycled to the process water after condensation or detoxified as mentioned above. No contaminated water is emitted to the air.

While most of the evaporated water is subsequently condensed, a small amount of the vapors

Figure 5. Detoxification facility for cyanides in a production plant

and noncondensables must be treated before release to the atmosphere. This is accomplished through the use of wet scrubbers, incineration, or a combination of both. Wet scrubber systems utilize liquid scrubbing solutions designed to contain or neutralize the noncondensable gases. Destruction efficiencies for these systems exceed 98 %. Incineration or thermal oxidation systems are also utilized for detoxification of the gasses. Destruction efficiencies for these systems exceed 99 %.

4. Quality Specifications, Impurities

The main impurities arising during the production of cyanides are cyanate (OCN^-), formate ($HCOO^-$), and carbonate (CO_3^{2-}). However, generally speaking, none of these are critical in any way for the applications in gold mining, the chemical industry, and electroplating. Very often these impurities are formed in application processes anyway.

Carbonates originate mainly from contact of excess alkali in the cyanide solution with atmospheric carbon dioxide. Hydrolysis of cyanide at higher temperatures (e.g., in drying sections, in hot solutions) leads to an increased content of formate, and ammonia is also formed. Cyanate can be formed by oxidation.

Ammonia can also occur as a possible impurity, especially in solutions after longer storage at elevated temperatures. If slow hydrolysis in solid (wet) material takes place, the ammonia formed can be released to the gas phase.

Due to reactions that lead to the formation of HCN and/or NH_3 commercial cyanide solutions can — mainly after prolonged storage under nonideal conditions — have a smell of HCN and/or NH_3. Absolutely dry cyanides normally do not smell of these species, as the presence of water is necessary for any of the decomposition reactions that lead to the release of HCN and/or NH_3.

For high-quality applications in the electroplating industry the content of a number of heavy metals is critical. Therefore, these must be below specified limits, which are, for example, defined in the DIN-ISO 50971. To achieve the DIN standards the respective alkali metal source (NaOH, KOH) must meet these requirements, as only these raw materials can act as sources of these metals. Due to its production processes, HCN is free of these metals.

For applications in high-quality electroplating and for most chemical synthesis applications, especially for the pharmaceutical industry, the specification for cyanides according to DIN-ISO 50971 can be regarded as a typical minimum quality requirement.

5. Uses

5.1. Gold Mining [27–32]

The majority of world sodium cyanide production is used in gold mining to extract gold from its ores (→ Gold, Gold Alloys, and Gold Compounds, Chap. 4.2). Elemental gold is very stable and does not react easily, even with very aggressive chemicals. Usually only in the presence of a strong oxidant together with a complexing agent can gold become water-soluble (formation of soluble gold salts). A classical example of how gold can be dissolved is the use of aqua regia, a mixture of the strong oxidant HNO_3 and the complexing agent HCl, where the chloride ion is the complexant for the gold. More weakly oxidizing conditions can be applied if a very strong complexing anion is used. In gold mining, cyanide is the complexing agent for the gold. Even by oxidation with oxygen (air, peroxides) under normal reaction conditions (dilute, slow reaction speed, moderate to low temperature) elemental gold is dissolved by oxidation and then subsequently complexed with cyanide:

$$4\,NaCN + 2\,Au + 1/2\,O_2 + H_2O \rightarrow 2\,Na\left[Au(CN)_2\right] + 2\,NaOH$$

While the gold chloride complex is unstable in dilute aqueous solution, even very low concentrations of the gold cyanide complex are stable in water over a wide pH range.

Cyanides are used in gold mining, although they are acutely toxic, mainly because of their high efficiency and selectivity, and they can not easily be substituted by other chemicals. In gold mining, not using cyanides would basically mean the termination of primary gold production. Only approx. 10 % of gold production today runs without the application of

cyanides. There have been numerous attempts to develop gold mining processes using alternative lixiviants for gold extraction. None of the alternative lixiviants or processes have ever been applied successfully in any large-scale operation on a worldwide basis. In this sense, only cyanide can fulfil the requirements of the gold mining industry to extract gold from the variety of ores in an economical and environmentally sound manner.

5.2. Chemical Synthesis

In a number of reactions, cyanides are used as starting materials or intermediates. Important examples in chemical synthesis are the cyanidation of acyl chlorides forming acyl cyanides as intermediate products:

$$RC(=O)Cl + NaCN \rightarrow RC(=O)CN + NaCl$$

and the nucleophilic substitution of halogenated alkanes by alkali metal cyanides for the elongation of the molecular chain by one carbon atom:

$$RBr + NaCN \rightarrow RCN + NaBr$$

Another basic reaction is the formation of cyanogen chloride from sodium cyanide and chlorine:

$$NaCN + Cl_2 \rightarrow ClCN + NaCl$$

Trimerization of cyanogen chloride results in the formation of cyanuric chloride (→ Cyanuric Acid and Cyanuric Chloride, Chap. 3), an important raw material for the production of herbicides, pesticides, fungicides, and insecticides. Cyanides are therefore important raw materials for crop-protection agents in the worldwide agricultural market. Many other organic synthetic procedures start with cyanogen chloride as raw material (→ Cyano Compounds, Inorganic, Chap. 4.4).

For a number of fertilizers and other important organic compounds production starts with cyanogen (→ Cyano Compounds, Inorganic, Chap. 5), the dinitrile of oxalic acid, which is made from alkali metal cyanide and copper sulfate, for example:

$$4\,KCN + 2\,CuSO_4 \rightarrow (CN)_2 + 2\,CuCN + K_2SO_4$$

In organic chemistry the cyano group basically acts as a building block which can easily be transformed into a variety of other chemical functional groups, such as carboxylic acids, esters, hydroxylamines, cyanohydrins, and many more. The standard application is to add cyanide to a molecule to modify its structure and functionality. The high-value end products are free of cyanide. The basic equations for the numerous reactions can be described as follows (X = halogen, R = organic group):

$$RX + CN^- \rightarrow RCN + X^-$$

$$RCN \rightarrow RCOOH \rightarrow RCOOR$$

Various inorganic synthetic products are made which still contain cyanide in an unmodified form. However these chemicals have specific characteristics such as colors (e.g., ferrous cyanide products) and do not exhibit the toxicity characteristics of free cyanides.

5.3. Electroplating of Metals

(→ Electrochemical and Chemical Deposition)

The electroplating industry uses silver and gold cyanides for electroplating processes:

$$KCN + AgNO_3 \rightarrow AgCN + KNO_3$$

$$KCN + AuCN \rightarrow K\left[Ag(CN)_2\right]$$

Beside the similar gold cyanide and potassium dicyanoaurate(I) the trivalent gold cyanide compound $K[Au(CN)_4]$ has the advantage of working in lower pH ranges and is able to deposit gold directly onto stainless steel base materials.

Electroplating of platinum metals is performed in cyanide baths and, for special applications, in cyanide melts. The high-temperature electroplating (HTE process) runs in the temperature range of 500 – 600 °C in an eutectic sodium – potassium cyanide mixture under argon gas.

The yellow and red cyanide complexes of bivalent and trivalent iron $K_4[Fe(CN)_6]$ and $K_3[Fe(CN)_6]$ are used to precipitate traces of base metals and to form blue pigments.

In the electroplating of metal surfaces complex metal cyanides are formed in solution or added to a process solution, and the metal from the cyanide complex is precipitated on the metal surface in an electric field. This technology is used to improve the surface characteristics of base metals and inexpensive metals (usually in industrial application).

5.4. Hardening of Metals

Strongly treated steel, such as that camshafts for motors, must be hardened by boron, nitrogen, or carbon. One of the most favorably applied methods is the carbonization of the metals in melts with cyanide. A superficial layer of iron carbide, Fe_3C (martensite) with a depth of several micrometers is formed at a temperature of about 900 °C. By the addition of barium chloride as activator the amount of cyanide could be minimized to some extent. The main chemical reactions of the hardening process are:

$$2\,NaCN + 2\,O_2 \rightarrow Na_2CO_3 + N_2 + CO$$

and

$$3\,Fe + 2CO \rightarrow Fe_3C + CO_2$$

and hence the overall reaction:

$$4\,NaCN + 4\,O_2 + 3\,Fe \rightarrow Fe_3C + 2\,Na_2CO_3 + 2\,N_2 + CO_2$$

The regeneration of such melts is possible with nonhazardous polymeric cyanic acid, azulminic acid $[CNH]_x$, cyanamide NH_2CN, dicyanamide, or melamine under nitrogen atmosphere.

6. Economic Aspects

The most important alkali metal cyanides in terms of production volume are sodium cyanide and potassium cyanide. Potassium cyanide production worldwide amounts to less than 10 % of the sodium cyanide production. The same production facilities can be used for the manufacture of NaCN and KCN.

In Europe, NaCN and KCN are produced at less than ten sites for various purposes. A significant amount is produced for captive use, but the majority is produced for applications within Europe and for export outside Europe.

Cyanide production efficiency is mainly driven by cost-effective production conditions. The production costs are mainly influenced by the costs for the raw materials HCN and NaOH/KOH, which can differ significantly.

In the 1990s the cyanide producer market in Europe consolidated to a significant extent, and a number of production sites have been closed. Existing production sites in Europe as of 2005 follow (asterisk denotes captive use):

Wesseling, Germany	NaCN solution, NaCN solid, KCN solid
Geleen, the Netherlands	NaCN solution
Billingham, United Kingdom	NaCN solution
Ludwigshafen, Germany	NaCN solution
Antwerp, Belgium	NaCN solution
Kolin, Czech Republic	NaCN solid
Pitesti, Rumania	NaCN solution
Seal Sands, United Kingdom*	NaCN solution
Saint-Avold, France*	NaCN solution
Roussillon, France*	NaCN solution

The production capacity of cyanides in Europe is in the range of 90 000 t/a (based on 100 % solid material). Most of this is solid sodium cyanide; a smaller part is solid potassium cyanide. Worldwide sodium cyanide production is estimated at approx. 500 000 t/a.

About 70 % of world cyanide production is used in the gold mining industry. The second biggest application is in industrial chemical synthesis, and third is electroplating together with metal hardening.

In Europe the majority of cyanide production is used in chemical synthesis. The second largest application is electroplating together with metal hardening, and the third is the mining industry.

The differences in prices between NaCN and KCN are mainly reflected in the price differences of the raw materials NaOH and KOH. As KOH is significantly higher in price, KCN is also more expensive than NaCN. Selling prices in Europe for NaCN/KCN are also subject to fluctuation on the world market. Prices were mainly less than 1650 \$/t for NaCN (100 %) and less than 3000 \$/t for KCN (100 %).

7. Best Available Techniques (BAT)

7.1. Processes

All above-mentioned techniques and procedures to produce cyanides (Chap. 2) in solution and in solid form only create a very small absolute contribution to emissions of hazardous chemicals to the environment. In addition, where properly designed wet scrubber systems are utilized to contain the noncondensed gases, there is no contribution to CO_2 emissions. Where incineration or combustion is used, there is an emission of CO_2. In production processes where wet solids are dried prior to compaction, the drying energy is typically provided by gas-fired heaters. The combustion gases from the heater are emitted to the atmosphere according to industry standards for fuel combustion devices. In Germany a typically permitted emission value for NO_x is approx. 1.8 kg/h for such heaters.

All of the above mentioned techniques can basically be considered to meet BAT requirements.

All melt processes must be considered to be very high in energy consumption and also indirectly contribute to CO_2 emissions due to the burning of fossil fuels. However, processes for HCN production, such as the BMA process, have high energy consumption for the formation of HCN, because the reaction of CH_4 and NH_3 (without oxygen) is endothermic and requires heating of the reaction gases to 1200 °C. The energy input necessary is 252 kJ/mol. In contrast the Andrussow process (using oxygen from air) is exothermic and creates energy in the range of 474 kJ/mol. However in the BMA process hydrogen (H_2) is formed, which can be used for the production of a number of downstream products such as H_2O_2, CH_3SH, and others. Also the energy content of H_2 must be taken into consideration positively if it is used to create energy for other processes. By doing this the energy balance of the Andrussow and the BMA process are very similar. Therefore, using HCN from the Andrussow or from the BMA process can be seen as equal regarding the overall energy balance.

Based on these considerations melt processes for the production of cyanides and the production of cyanides based on HCN derived from the BMA process can be judged to be similar with respect to energy consumption and the respective CO_2 release in case organic fuels are burned.

Finally, when cyanides are detoxified by an oxidation process cyanate (OCN^-) is formed and this can be broken down in aqueous solution to ammonium carbonate $(NH_4)_2CO_3$ (depending on the pH of the solution). From this solution ammonia can be released already at slightly alkaline pH (pH equilibrium curve for the system NH_3/NH_4^+ is relevant here). In the end this means that the NH_3 used for the synthesis of HCN is again released as NH_3. The carbon from the hydrocarbons (CH_4) is transferred into carbonates.

Any process leading directly to a shippable solid material which meets the quality and safety standards of the customer industry would be ideal to be considered as BAT. This is the case for the patented process to form fluidized-bed granulates, but this process is not in large-scale operation yet.

As a final conclusion all existing processes for cyanide production can be considered as BAT at present (2005).

7.2. Procedures

Destruction of Cyanides in Waste Gases and Wastewaters (Fig. 6)**.** Due to their toxicity, cyanides are removed from an air stream (3400 m^3/h, 50 mg/m^3 HCN) by using a scrubber with NaOH solution. The solution is circulated between a buffer vessel and the washer. The solution is sampled regularly and exchanged if the content of free OH^- is too low for the absorption of HCN from the air stream. The cyanide solution is then reconditioned with other CN^--rich wastewater streams and used to substitute raw materials (1 m^3/h, 5 % CN^-). A

low-CN^- wastewater stream (15000 m^3/h, 700 mg/m^3 HCN) leaves the reconditioning process. The remaining cyanide is destroyed together with low-CN^- wastewater streams by pH adjustment and oxidative destruction with H_2O_2. The overall efficency is 99.9 %. Detoxification of cyanide at the plant with hydrogen peroxide based technology should be considered as BAT as no salts are produced and the only byproduct is water.

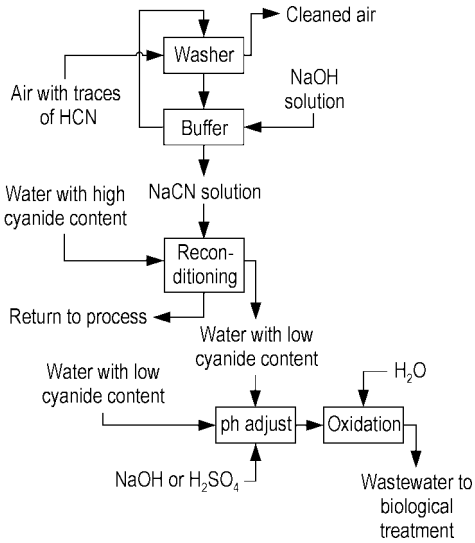

Figure 6. Flow chart of the cyanide destruction in waste gases and wastewaters

Environmental Benefits.
- Removal of HCN/CN^- from waste gases and wastewaters
- Achieved emission values for waste gas (concentration 1 mg/m^3, mass flow 3 g/h) and wastewater (concentration 1.1 mg/m^3, mass flow 2.3 g/h)

Cross-Media Effects.
- Reuse of waste streams as raw materials
- No AOX in wastewater
- No contaminated waste
- Lower energy consumption

Applicability.
- Generally applicable

Economics.
- Cost reduction for wastewater treatment

Driving Force for Implementation.
- Removal of toxic compounds from waste gases and wastewaters

Incineration of Process Air Containing Volatile Organics. A raw material for the process can contain small amounts of volatile organics. In the process, benzene was found in several air streams (2500 m^3/h, variable benzene concentration), which are sent to an air cleaning system. The air streams which contain benzene are identified and separated from the others. The benzene containing air streams are burned in a air heater which is run by natural gas (Fig. 7). The energy from the incineration is recovered. The circulation drying air is heated with this energy. The overall efficiency is 99.9 %.

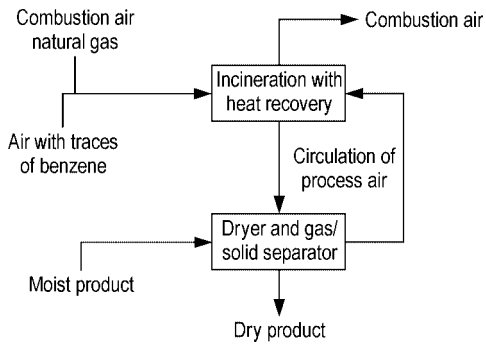

Figure 7. Incineration of process air containing volatile organics

Environmental Benefits.
- Removal of benzene from waste gases
- Achieved emission values for waste gas (concentration <1 mg/m^3, mass flow 0.4 g/h)

Cross-Media Effects.
- Reduced consumption of natural gas by using benzene-containing air
- Heat from incineration is recovered in the process

Applicability.
- Generally applicable

Economics.
- Reduction of energy costs
- Comparatively low investment costs

Driving Force for Implementation.
- Removal of toxic compounds from waste gases

Cleaning of Equipment Contaminated with Cyanides in Totally Enclosed Systems. During production, solid cyanide builds up in pipelines, machines, and vessels and can thus cause a blockage. In all equipment, a cleaning-in-place (CIP) system was installed (Fig. 8). In case of a blockage, the equipment can be easily rinsed and cleaned in a totally closed system. No contaminated emissions to the environment or exposure of workers occurs. The rinsing water containing cyanide (15 m^3/h, variable cyanide concentration) is collected in a closed piping system and stored in vessels. The rinsing water is then recycled into the process to reduce the use of raw material. In addition, solid waste containing solid cyanide salts (i.e., off-spec product, waste) can be dissolved in the collecting vessels and recycled in the process. Thus the amount of hazardous wastes is reduced. The overall efficiency is 99 %.

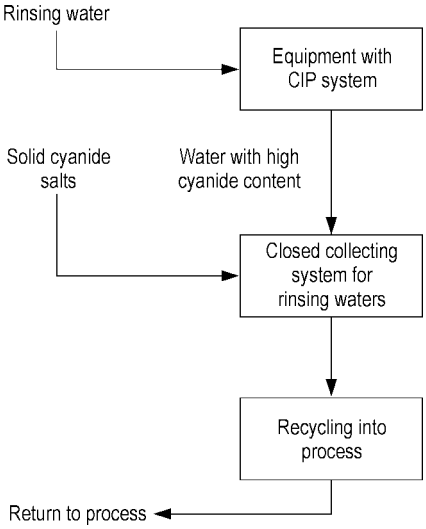

Figure 8. Cleaning of equipment contaminated with cyanides in totally enclosed systems

Environmental Benefits.
- No solid waste containing cyanides
- Recycling of solid waste within the process
- No emissions during cleaning

Cross-Media Effects.
- Reduced amount of rinsing water through CIP system
- Lower energy consumption

Applicability.
- Generally applicable

Economics.
- Cost reduction for wastewater treatment
- Cost reduction for raw materials

Driving Force for Implementation.
- Reduction of emissions to environment and workplace

Removal of HCN and NH_3 from Waste Gases. Due to their toxicity, HCN and ammonia are removed from gas streams (3400 m^3/h, 50 mg/m^3 HCN, 250 mg/m^3 NH_3) using an alkaline scrubber with NaOH solution and an acidic scrubber with sulfuric acid. For each scrubber the solution is circulated between a buffer vessel and the washer. The solution of each washer is sampled regularly and exchanged if the content of free OH^- is too low for the absorption of HCN from the air stream or if the free H_2SO_4 concentration is too low for the absorption of ammonia. The cyanide solution of the alkaline washer is recycled to the process to substitute raw materials. The ammonium sulfate solution of the acidic washer is recycled in a nearby plant to produce fertilizer. The secondary containment of the acidic washer is strictly separated from the secondary containment of the alkaline washer and the rest of the production building (Fig. 9). If acids came into contact with solutions containing cyanide, HCN would be emitted to the environment and workplace. The overall efficiency is 99.9 %.

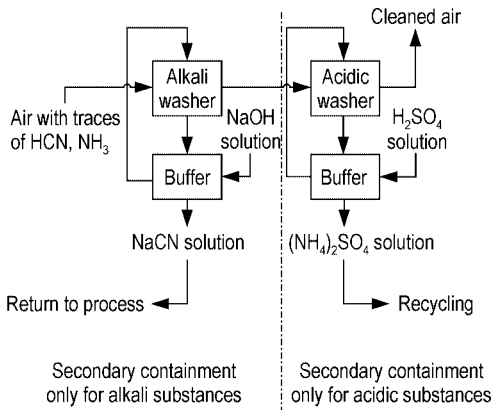

Figure 9. Removal of HCN and NH_3 from waste gases

Environmental Benefits.
- Removal of HCN/NH_3 from waste gases
- Strict separation of acidic and alkaline containment reduces emissions in case of an emergency
- Achieved emission values for HCN (concentration 1 mg/m^3, mass flow 3 g/h) and NH_3 (concentration 1.2 mg/m^3, mass flow 3 g/h)

Cross-Media Effects.
- Reuse of washer solutions as raw materials
- No wastewater from scrubbers
- Lower energy consumption
- Simple and reliable control of efficiency by regular analysis of the solution

Applicability.
- Generally applicable

Economics.
- Low operating costs

Driving Force for Implementation.
- Removal of toxic compounds form waste gases

Returnable Packages for Solid Cyanides. Besides one-way packages (i.e., steel drums, bulk bags) returnable packages are used for the transport of cyanides. In a filling line, the packages are filled with cyanides. After the customer has emptied the packages, the sealed contaminated packages are returned. A second option is that the part of the package that did not come into contact with cyanide is disassembled and returned to the filling line. Here it is checked and assembled again. Cyanide customers can be equipped with unloading devices to unload the packages in a closed system.

Environmental Benefits.
- Reduction of cyanide dust emissions during unloading
- Reduction of the amount of contaminated packages and waste

Cross-Media Effects.
- Higher safety standards in workplace during unloading
- Return transport of packaging to the filling station leads to additional fuel consumption

Applicability.
- Generally applicable

Economics.
- Reduction of packaging costs

Driving Force for Implementation.
- Reduction of contaminated waste

Transport of Solid Cyanide. Besides one-way packages (i.e., steel drums, bulk bags) returnable packages are used for the transport of cyanides. In a filling line, the vessels that are normally used for a liquid transport are filled with solid cyanides. For the unloading the cyanide customer connects two hoses to the vessel. The solid cyanide is dissolved by circulating water through the vessel, which has built-in nozzles. The cyanide is unloaded in a closed system.

Environmental Benefits.
- Reduction of cyanide dust emissions during unloading
- Reduction of the amount of contaminated packages and waste

Cross-Media Effects.
- Higher safety standards in the workplace
- Greater safety through the transport of solid material instead of liquid solution
- Return transport to the filling station leads to additional fuel consumption

Applicability.
- Generally applicable when cyanide is dissolved in water

Economics.
- Reduction of packaging costs

Driving Force for Implementation.
- Reduction of contaminated waste

Operation of Sodium Cyanide Plant. The operation of a sodium cyanide plant requires numerous and frequent observations of operating parameters, and adjustment of the process controls to ensure safety and environmental control. A computer process control system can monitor hundreds of process parameters many times per minute, and automatically adjust parmameters to ensure a safe and environmentally sound operation. Using computerized process control, automatic interlocks can be implemented to prevent

overfilling tanks, emitting excess contaminants to the atmosphere, and minimize the generation of wastewater.

Environmental Benefits.
- Reduction of cyanide emission
- Reduction of contaminated wastewater

Cross-Media Effects.
- Higher safety standards in workplace
- Lower emissions in air and water streams
- Reduced need to "open" the process for cleaning

Applicability.
- Generally applicable for solid sodium cyanide production facility

Economics.
- Reduction of waste treatment

Driving Force for Implementation.
- Reduction of emissions

Double Containment of Solid Sodium Cyanide Plant. In a solid sodium cyanide production facility, sometimes there are leaks, dust, equipment wash water, etc. which can result in sodium cyanide solution being present on the floor of the production plant. Plants constructed of concrete or other traditional civil construction materials are subject to potential cracking. Through these cracks, the cyanide-contaminated water can leach into the ground and possibly groundwater. Providing double containment for the ground floor and sumps reduces the potential for cyanide to be released to the soil or groundwater.

Environmental Benefits.
- Reduction of soil contamination
- Reduction of contaminated civil construction materials

Cross-Media Effects.
- Higher safety standards in workplace
- Greater safety of soil and groundwater

Applicability.
- Generally applicable when cyanide solutions are present on the ground floor

Economics.
- Reduction of material losses

Driving Force for Implementation.
- Improved assurance of no leakage of cyanide solution to the soil

Storage of Solid Alkali Metal Cyanide. Sodium cyanide is stored in UN-certified packages for the transport of dangerous goods in specially designed storage buildings. The building is fenced and locked to prohibit the access of unauthorized persons. Alkali metal cyanide easily dissolves in water. Therefore the building is constructed such that no rainfall or surface water can enter the building. Additionally, the use of water is prohibited for fire fighting. Alkaline foam is used instead. The use of acidic foams is also prohibited since HCN can be liberated by this kind of foam.

Environmental Benefits.
- Reduction of soil and water contamination
- Reduction of contaminated civil construction material

Cross-Media Effects.
- Higher safety standards in workplace
- Higher security standards to limit access to the cyanides

Applicability.
- Generally applicable where cyanide salts are stored in packages

Economics.
- Comparatively low construction costs

Driving Force for Implementation.
- Enhancing safety and security standards

Training of Personnel in a Cyanide Plant. To operate a cyanide plant a high level of knowledge and training is necessary. A complex process, modern technologies, highly toxic substances, and emergency situations place a lot of stress on the operators. To ensure that incidents and accidents are kept to a minimum and to reduce mistakes in emergency situations to a minimum special education and training programs are necessary. Besides a sound basic education in chemical operations, personnel are trained in different level of jobs over the years. The skills are tested in written or practical exams. Additionally, all operators are trained regularly in emergency response, occupational, product, and transportation safety regulations.

Environmental Benefits.
- Reduced risk of emission to air, water, and soil in case of emergencies

Cross-Media Effects.
- Higher education level of personnel

Applicability.
- Generally applicable

Economics.
- Better education leads also to a reduction in production costs

Driving Force for Implementation.
- Enhancing safety standards

7.3. Toxicology and Occupational Health

For toxicology and occupational health of NaCN, KCN, and HCN, see → Cyano Compounds, Inorganic, Chap. 6.

8. References

General References

1. CyPlus, More than Cyanide (Planning, Research & Development, Production, Packaging, Transport, Storage, Handling & Use, Disposal & Recycling), company brochure, *Frankfurt am Main*, Germany, 2003.
2. Degussa, Stets geforscht. . ., company brochure, Frankfurt am Main, Germany, 1988.
3. Degussa, *Edelmetalltaschenbuch*, Hüthig-Verlag, Heidelberg, Germany, 1995.
4. Degussa, Metallforschung und Entwicklung, company brochure, *Frankfurt am Main*, Germany, 1991.

Specific References

5. R. Norcross, N. Steiner, S. Gos, A. Rubo: "The Cyanide Life Cycle — Packaging & Transport", *Mining J. Mining Environ. Manag.* 2001, May, 19 – 21.
6. CEFIC, Cyanide Sector Group — Guidelines for Storage, Handling, and Distribution of Alkali Cyanides, May 2000, under revision in 2004.
7. VCI Brochure "Responsible Care", (Umweltschutz, Produktverantwortung, Arbeitssicherheit & Gesundheitsschutz, Anlagensicherheit & Gefahrenabwehr, Logistik, Dialog), Frankfurt am Main, Germany, July 2003.
8. S. Gos, A. Rubo: "Responsible Application of Cyanides", Randol Gold & Silver Forum '97, Monterey, California, May 18 – 21, 1997, pp. 161 – 165.
9. A. Rubo, S. Gos, A. Dickmann, A. Dominguez-Perez, R. Norcross: "Effects of Responsible Care on the Risks, Legal and Cost Aspects of Cyanides", Randol Gold & Silver Forum '99, Denver, Colorado, May 11 – 14, 1999, pp. 53 – 56.
10. A. Dominguez, A. Rubo, S. Gos: "Relation between Risks, Legal Aspects and Costs in the Application of Cyanides", International Mining and Environment Congress "Clean Technology: Third Millenium Challenge", Lima, Peru, July 12 – 16, 1999.
11. A. Rubo, S. Gos, A. Dickmann, A. Dominguez Perez: "Risk Related Cost Parameters in Handling Cyanides", 4th International Gold Symposium, Lima, Peru, May 02 – 05, 2000.
12. N. Steiner, A. Rubo, A. Aivasidis: "Responsible Cyanide Management in the Gold Mining Industry" Exploitation of Gold Deposits in Thrace, International Meeting, Komotini, Greece, Oct. 14 – 15, 2000.
13. S. Gos, A. Rubo, A. Dominguez-Perez: "La Relevancia De Lixivantes Alternativos EN Conexion Con Aspectos Technicos, Seguridad Ocupaccional Y Seguridad Ambiental", "Technical Factors, Occupational Safety And Environmental Concerns", Argentinia, Panorama Minero, No. 255, Dic. 2000, pp. 48 – 50.
14. F. Harenburg and A. Rubo: "Responsible Care, Sustainable Development in Gold Mining and the Future Role of CyPlus", Johannesbourg, South Africa, Indaba, 2003.
15. ISO 9001:2000/14001 and ICMC certification.
16. CyPlus, press release Jan. 2004 on Certification on ICMC, ISO 9001:2000 and ISO 14001, Hanau-Wolfgang, Germany.
17. CEFIC, press release on "Alkali Cyanide Mutual Aid Scheme", Brussels, Belgium, Oct. 2000.
18. CEFIC, "Responsible Care Status Report", Brussels, Belgium, 2002.
19. CEFIC, "Managing Chemicals. Voluntary and Obligatory Programmes", Brussels, Belgium, 2003.

20. A. Smith and T. Mudder: The Chemistry and Treatment of Cyanidation Wastes, *Mining Journal Books Limited*, London 1991.
21. A. Rubo, A. Dickmann, S. Gos: "Laboratory Simulation of HCN Emissions from Tailings Ponds", Tailings and Mine Waste '00, 2000 Balkema, Rotterdam, pp. 307 – 313, Proceedings of the 7th International Conference on Tailings and Mine Waste '00, Fort Collins, Colorada, USA, 23 – 26 Jan. 2000.
22. CyPlus CombinOx^{TM}, A New Technology For Cyanide Destruction, E.A. Devuyst/Inco Tech, Johannesbourg, South Africa, Indaba 2003, presentation on request through CyPlus.
23. J. Lorösch, Process and Environmental Chemistry of Cyanidation, Degussa, Frankfurt am Main, Germany, 2001, pp. 401 – 406.
24. A. Griffiths, H. Knorre, S. Gos, R. Higgins: "The Detoxification of Gold-Mill Tailings with Hydrogen Peroxide", *J. S. Afr. Inst. Min. Metall.* **87** (1987) no. 9, 279 – 283.
25. T. Mudder, S. Whitlock: "Biological Treatment of Cyanidation Wastewaters", *Min. Metall. Process.* (1989) 161 – 165.
26. A. Rubo, S. Gos: "Treatment after Cyanide Contamination", *Mining Magazine* 2002 240 – 246.
27. A. Rubo, Degussa-Huels AG, "The Cyanide Life Cycle", Workshop "Management of Cyanide in the Gold Industry Beyond 2001", AJ Parker Cooperative Research Centre for Hydrometallurgy, Dec. 5, 2001, Kalgoorlie, Australia.
28. Golden Prospects in Europe, brochure, published by euromines, Avenue de Broqueville, 12, B-1150 Bruxelles, 2002, sponsored by Boliden, Newmont, Normandy Madencilik A.S., Rio Narcea Gold Mines LTD, Thracean Gold Mining S.A., CyPlus GmbH (Degussa Group).
29. Euromines brochure, Occupational Health and Safety Management in the European Mining Industry, Brussels, Belgium.
30. E. von der Linden: "Goldbergbau und Zyanlaugung - Die Rolle der Banken in der Finanzierung", *Erzmetall* **7/8** (2000) 471 ff.
31. M. Ericsson: "M& A im Bergbau erreichen im Jahr 2001 einen neuen Höchststand, Gold und Eisenerzbau", *Erzmetall* **3** (2002).
32. S. Gos, A. Rubo, Degussa-Huels AG, "Alternative Lixiviants for Gold Leaching — a Comparison", Randol Gold & Silver Forum 2000, Vancouver, BC, Canada, April 25 – 28, 2000, pp. 271 – 281.

Computational Fluid Dynamics

ANJA R. PASCHEDAG, University of Technology Berlin, Berlin, Germany (Chap. 1)

1. Introduction 2
2. Procedure 2
3. Modeling 3
3.1. Transport Equations 3
3.2. Initial and Boundary Conditions . . . 5
3.3. Turbulent Flow 6
3.4. Multiphase Approaches 9
4. Numerics 12
4.1. Basics 12
4.2. Finite Volume Method 14
4.3. Pressure Correction Methods 17
4.4. Lattice Boltzmann Method 17
5. Interpretation 18
6. Industrial Application 19
7. References 21

Computational Fluid Dynamics (CFD) is a numerically based tool for the prediction of flow field, concentration and temperature distribution. Its main parts are mathematical modeling, discretization, numerical solution of the discretized equations and the interpretation of numerical results.Basic equations in all mathematical models for CFD are balances for momentum and total mass determining velocity, pressure and density field. Depending on the case considered they are supplemented by mass balances for single species and a heat balance. Additional models are required to describe, e.g. turbulence, multiphase flows, chemically reactive systems and other special cases.Basis for the discretization of the balance equation is the discretization of a space — the grid generation. Most codes can handle unstructured grids. Nevertheless, certain requirements concerning grid structure have to be fulfilled to get stable convergence and an accurate solution. Traditionally, most CFD codes use finite volume discretization for the balance equations, even if finite element algorithms are of increasing relevance for simulations with adaptively moving grids and for coupling CFD with structural dynamics simulations. A new approach for simulations with high resolution in space and time is the lattice Boltzmann method.

Finally, the numerical results have to be graphically presented and interpreted. Because of the huge amount of numerical data provided by each simulation this cannot be done with a general method. It must always be analysed in reference to a certain research question. Errors caused by the model formulation and by the numerical scheme have to be analyzed in order to judge the accuracy of a simulation. Quantitative estimates are required for an adequate interpretation of the results.

Symbols

A	m^2	cell surface (FVM)
$C_{BC1} \ldots C_{BC4}$		constants of mathematical expressions for boundary conditions
$C_\mu, C_{\varepsilon 1}, C_{\varepsilon 2}$		constants in k—ε model
c	mol/m^3	concentration
c_p	$J\,kg^{-1}K^{-1}$	specific heat capacity
C_Δ		Courant number
D	m^2/s	diffusion coefficient
$\mathbf{F}$	N	force
$\mathbf{F}$		general function
f		weighting factor in time discretization
G		filter function (LES)
$\mathbf{g}$	m/s^2	gravitational acceleration
I	$kg{\cdot}m^2$	moment of inertia
J_ϕ		molecular flux
k	m^2/s^2	turbulent kinetic energy
m	kg	mass
$\mathbf{n}$	m	normal vector
N		number of nodes
$\hat{P}$	m^2/s^3	source term in transport equation for k
p	N/m^2	pressure
$\dot{q}$	$J\,m^{-2}s^{-1}$	heat flux density
r		residual
Re		Reynolds number
S_c		source in concentration equation
S_{hr}		source in temperature equation
S_u		source in energy equation
S_ϕ		source in general transport equation
$\mathbf{T}$	N·m	torque
T	K	temperature
t	s	time
u	J/m^3	specific internal energy

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.i07_i01

V	m^3	volume
$\mathbf{v}$	m/s	velocity
$\mathbf{x}$	m	position vector
α_i		volume fraction of phase i
Γ	m^2/s	molecular transport coefficient
δ		unit tensor
ε	m^2/s^3	energy dissipation rate
κ	Pa·s	dilatational viscosity
λ	J ($m^{-1}s^{-1}K^{-1}$)	heat conductivity
μ	Pa·s	dynamic viscosity
μ_t	Pa·s	turbulent viscosity (RANS)
ρ	kg/m^3	density
$\boldsymbol{\tau}$	N/m^2	Newtonian shear stress
$\boldsymbol{\tau}^s$	N/m^2	turbulent stress tensor (LES)
τ_t	N/m^2	turbulent stress tensor (RANS)
ϕ		general transport quantity
Ω	m/s	rotational speed of particles
Indices		
B,E,N,P,S,T,W		nodes of the FVM grid
b,e,n,s,t,w		cell faces of the FVM grid
Pi		i^{th} particle parcel

1. Introduction

The increasing pressure on development time and product quality in industrial process development stimulates the application of numerical methods in addition to traditional experimental ones. In many cases the numerical tools can save time and costs if they are used complementary to experimental methods. The simulation tool for modeling fluid dynamic processes is called computational fluid dynamics (CFD). Highly complex mathematical models are needed for a sufficient description of the relevant interactions in reactive flows occurring on different timescales. Even if not all aspects can be considered in detail to keep the equation system in a manageable frame, the complex mathematical models are far from being solvable analytically. Therefore, numerical methods play an important role in CFD, and many users associate CFD mainly with its numerical aspects. But the subject is defined much more widely: formulation of the mathematical model, visualization, and interpretation are all essential parts of the tool.

2. Procedure

This chapter sketches the general steps of a CFD simulation. It gives an overview of the procedure and the relation of the individual parts. The points mentioned here are described in more detail in the following chapters.

As with all investigation methods, the formulation of the exact question is a crucial point for the entire CFD procedure. The question includes the target value and its accuracy, the dependencies of importance and possible approximations and assumptions which can be made. Also the range in space and time to be considered is a part of it. All significant physical parameters or the range of their variation must be specified together with the physical state at the boundaries and at the initial time. The formulation of the question determines the effort and the time required to solve the problem and should be therefore as strict as necessary and as weak as possible.

The mathematical model can be formulated on the basis of the question posed. Its central elements are balance equations for mass, momentum and energy. These are supplemented by additional model equations if necessary. All partial differential equations require initial and boundary conditions for their solution. Parameters and constants in the equations have to be specified in a way that the model is mathematically closed. For this purpose, additional laws, like an equation of state or temperature dependencies of physical properties (e.g., density, viscosity) may be required. Not always can the most exact mathematical formulation be solved with reasonable effort. In such cases, additional model assumptions such as averaging procedures, reduction of the number of variables or restriction of the range in space and time considered must be introduced.

The resulting system of differential, integral and algebraic equations is too complex to be solved analytically. Therefore, numerical algorithms are applied to derive a system of linear algebraic equations which provides an approximative solution of the original mathematical model. In most cases this discretization is performed by using the finite volume method or the finite element method. Specialized methods exist for certain applications. A relatively new alternative to classic discretization algorithms are statistical methods like the lattice Boltzmann method.

As a result of the great complexity of the mathematical model, the numerical solution of the governing equations must be postprocessed to extract the information necessary to answer

the original question. Most often this information can best be presented in a graphical plot, but the data to be plotted are usually not the direct solution of the model. Relevant data must be selected, linked, transformed, and in some cases statistically treated and normalized to create a data set which allows the user to answer the question.

Last but not least, the accuracy of the numerical result must be judged, even if the simulation is not erroneous in an obvious sense. Starting from the setup of the mathematical model over all numerical steps, up to postprocessing, inaccuracies are introduced or at least accepted. Only a rough quantitative error estimation can prove whether the solution provided is accurate enough for the question considered.

3. Modeling

The central CFD modeling approach is to balance momentum, mass and heat. This approach is based on the continuity hypothesis which states that all balance quantities and physical parameters change continuously within each phase. Consequently, limiting considerations for infinitesimally small volumes never consider changes at the molecular level. Discontinuous jumps are found only at interfaces. Therefore, in some points, modeling of multiphase systems requires different approaches than modeling of single-phase systems.

For an easier understanding of the following discussion a few terms will be defined:

- Balance quantities: Quantities governed by laws of conservation can be balanced. In general, mass, momentum and energy are such balance quantities. For detailed considerations masses of different chemical species or different kinds of energy can be balanced separately. In such balances source terms for possible conversions appear. Balance quantities are extensive values.
- Transport quantities: While the basic step of balancing is an integral consideration of the balance quantities, differential transport equations can be derived from the original balances. Transport equations describe the distribution and transport of intensive values, either mass-based (mass fraction, specific energy) or volume-based (density, concentration), which are the so-called transport quantities.
- Independent variables: Independent variables in transport equations are normally the three spacial directions and time. In special cases symmetry can be used to reduce the number of spacial directions; in the case of stationary considerations, time is not considered. For certain model approaches additional independent variables must be included. In probability density models, e.g. for micromixing, the species composition is an additional independent variable, in population balances particle properties such as the characteristic diameter take on this functionality.
- Parameters: All known values in the transport equations — besides transport quantities and independent variables — are called parameters. "Known" must be understood in a mathematical sense: the solution of the equations is only possible if values for the parameters are given. In the physical sense the determination of these values might be problematic or only possible with the help of an additional equation.

3.1. Transport Equations

Transport equations are derived on the basis of balances. For the general transport quantity ϕ the transport equation reads:

$$\begin{array}{ccccccc} \dfrac{\partial \phi}{\partial t} & = & -\nabla(\mathbf{v}\phi) & - & \nabla \mathbf{J}_\phi & + & S_\phi \\ (a) & = & (b) & + & (c) & + & (d) \end{array} \tag{1}$$

Term (a), the time derivative, codes transient changes. Term (b) models the convective transport with $\mathbf{v}$ being the convective velocity. Term (c) in this general form is the molecular transport term including the molecular flux vector $\mathbf{J}_\phi$. In certain cases, in particular in turbulence modeling, structurally similar terms can occur which are numerically treated in the same way but have different physical meanings. Finally, term (d) is the source term. The kinds of sources which occur depend strongly on the transport quantity ϕ. In the momentum balance these are forces, in the energy balance these are exchange rates to other types of energy, and in mass balances for single species these are conversion rates due to chemical reactions.

Detailed derivations and discussions of the transport equations can be found in [1] (see also Transport Phenomena). Only a short overview is presented here.

Continuity Equation. A basic physical law is the conservation of mass. From this the continuity equation can be derived:

$$\frac{\partial \rho}{\partial t} = -\nabla\ (\rho \mathbf{v}) \tag{2}$$

As there is no molecular mass transfer relative to the convective velocity (with its mass-based definition) and no source of mass in a nonrelativistic system, this equation contains only terms (a) and (b) of Equation 1 with density ρ being the transport value.

Equation of Motion. CFD considers transport phenomena in fluids, so the convective velocity is a central quantity. This becomes obvious also from Equation 1, where velocity is the only general, ϕ-independent value. The velocity is determined by the equation of motion:

$$\frac{\partial \rho \mathbf{v}}{\partial t} = -\nabla\ (\rho \mathbf{v}\mathbf{v}) - \nabla\, \boldsymbol{\tau} - \nabla p + \rho \mathbf{g} \tag{3}$$

The velocity itself is not the transport quantity here but rather the volume-based momentum $\rho \boldsymbol{v}$. The stress tensor $\boldsymbol{\tau}$ plays the part of the molecular flux. It is a function of the velocity gradient, but this functionality can not be expressed in a general form because it depends on the rheologic properties of the fluid. In many applications a Newtonian fluid can be assumed, and τ can be expressed in the following way:

$$\boldsymbol{\tau} = \mu[\nabla \mathbf{v} + (\nabla \mathbf{v})^{\mathbf{T}}] + \left(\kappa - \frac{2}{3}\mu\right)(\nabla \cdot \mathbf{v})\delta \tag{4}$$

where μ is the viscosity, which is independent of shear stress and time in Newtonian fluids, and κ is the dilatational viscosity, which in most real cases has no influence on the flow.

The last two terms of Equation 3 are the source terms to be considered in most cases. The pressure term is of greatest importance. It has a dual role in the structure of the equation. As mentioned, it can be seen as a source term, but it also could be counted as a special type of stress and included together with τ in a general stress term. The first concept is applied here because it provides a more suitable structure for the discussion of numerical methods.

Gravity acts in all systems and is therefore also written here. In special cases other forces acting on a fluid like electromagnetic forces or centrifugal forces also must be considered.

Two things complicate the numerical handling of the equation of motion. The first is the nonlinear structure of the convective terms. While in other transport equations the velocity can be handled like a parameter, in the equation of motion it is the unknown. The second problem results from the pressure term. There is no transport equation available for pressure. In compressible fluids an equation of state is used additionally which allows one to calculate the pressure if the density field is known from the equation of continuity. The solution of these three coupled equations (one of which has three components) is numerically not handsome, but possible without significant problems. In incompressible systems the density is constant and an equation of state cannot be used to determine the pressure. The remaining equation for the pressure is the equation of continuity, but it does not contain the pressure. The relation can be described such that only the velocities which satisfy the equation of motion with the right pressure gradient also satisfy the equation of continuity.

Navier–Stokes Equations. A special and commonly used version of the equation of continuity and the equation of motion are obtained for Newtonian fluids with constant density and constant viscosity. The equations simplify to:

Equation of continuity:

$$\nabla \mathbf{v} = 0 \tag{5}$$

Equation of motion:

$$\rho \frac{\partial \mathbf{v}}{\partial t} = -\rho \mathbf{v} \nabla \mathbf{v} + \mu \nabla^2 \mathbf{v} - \nabla p + \rho \mathbf{g} \tag{6}$$

This equations are called Navier–Stokes equations. They are of restricted validity because changes in temperature as well as changes in the chemical composition effect density and velocity. Nevertheless, they are a reasonable approach in many cases.

Concentration Equation. The equation of continuity results from the overall mass balance. If a multicomponent system is considered and

the concentration of the different species is of interest for the simulation, separate transport equations must be solved for them. As the global mass balance is solved in all cases, the number of additional equations needed for a complete description is one less than the total number of species in the system. The transport equation for a concentration c reads:

$$\frac{\partial c}{\partial t} = -\mathbf{v}\nabla c + \nabla(D\nabla \mathrm{c}) + S_{\mathrm{c}} \tag{7}$$

where D is the molecular diffusion coefficient and S_{c} is the source term resulting from a chemical reaction. It is expressed via the chemical reaction kinetics, which in many cases include nonlinear expressions and dependencies between concentrations of different chemical species.

Energy Equation. Formulating energy equations is a large area because there are many different forms of energy which can be converted into each other. For CFD the equation of the internal energy is usually considered to obtain an equation for the temperature profile. For the application of turbulence modeling, a certain part of the kinetic energy is of interest. This second case is discussed in Section 3.3 only the internal energy is considered here. Internal energy covers the kinetic energy of atomic and molecular motion, potential energy of intermolecular interactions, energy of chemical bonds, and nuclear energy. Not all aspects of this energy can be determined absolutely. Therefore, only differences in the internal energy between two states are discussed. Balancing of the internal energy gives:

$$\begin{aligned} \frac{\partial \rho u}{\partial t} &= -\nabla(\rho u \mathbf{v}) - \nabla \dot{\mathbf{q}} - (\boldsymbol{\tau}:\nabla \mathbf{v}) - p(\nabla \cdot \mathbf{v}) + S_{\mathrm{u}} \\ (a) &= (b) \qquad + (c) \quad + (d) \qquad + (e) \qquad + (f) \end{aligned} \tag{8}$$

with

(a) = temporal change
(b) = convective transport
(c) = heat conduction
(d) = increase by viscous dissipation
(e) = increase by compression
(f) = conversion from other types of energy

If an incompressible fluid is considered, in addition viscous dissipation, radiation and other kinds of energy conversion can be neglected, and if the mixing enthalpy is also negligible, the only potential source is the enthalpy of a chemical reaction S_{hr}. Density, thermal conductivity, and heat capacity values are calculated for the mixture in the case of a multicomponent system. With these assumptions the temperature equation can be derived in the form of a commonly used equation:

$$\frac{\partial \rho c_p T}{\partial t} = -\mathbf{v}\nabla(\rho c_p T) - \nabla(\lambda \nabla \rho c_p T) + S_{hr} \tag{9}$$

where c_{p} is the heat capacity and λ the thermal conductivity. If the assumptions do not hold as described above, extended forms of the temperature equations must be derived.

3.2. Initial and Boundary Conditions

Partial differential equations cannot be solved without initial and boundary conditions. In a mathematical sense this can be interpreted such that the solving of a differential equation requires its integration, and to perform this the integration constants must be determined. Consequently, the number of additional conditions needed for each independent variable is equal to the number of derivatives. Therefore, in time only one condition is needed, which in most cases is given at the initial time, while for the space coordinates two conditions are required, which are realized normally by one condition at each boundary point.

For the initial conditions values of the transport quantities ϕ must always be given. These values may vary in space but cannot be a function of ϕ.

Three different types of boundary conditions are possible:

Dirichlet condition	
$\phi\|_{\mathrm{BC}} = C_{\mathrm{BC1}}$	fixed value of ϕ given at the boundary
Neumann condition	
$\frac{\partial \phi}{\partial n}\|_{\mathrm{BC}} = C_{\mathrm{BC2}}$	derivative of ϕ normal to the boundary is given
Cauchy condition	
$\frac{\partial \phi}{\partial n}\|_{\mathrm{BC}} + C_{\mathrm{BC3}}\,\phi\|_{\mathrm{BC}} = C_{\mathrm{BC4}}$	combination of value and derivative of ϕ is given

In a finite region of the boundary a Dirichlet condition is required. The use of these boundary

conditions can be segregated in relation to the physical nature of boundaries. The ones most often used are presented here, but for special flow situations other types can also be defined. It is not always easy to find the expression required, and in some cases a setup is more easily obtained by moving the boundary location than by constructing a complex formulation at the original location. An example is shown in Figure 1.

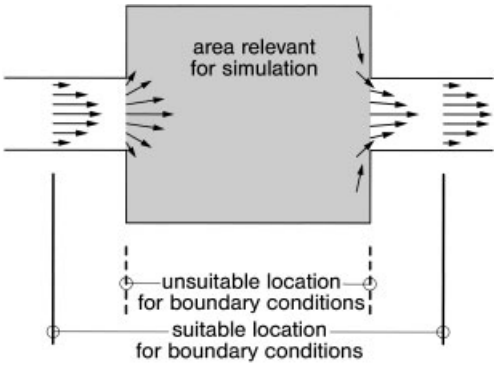

Figure 1. Location of boundary conditions

- Inlet: It is assumed that the conditions at the inlet are known. Therefore, fixed values are given at the inlet for all transport quantities. These fixed values may vary along the inlet, e.g., a parabolic velocity profile can be given for a laminar flow in a tube.
- Wall: Most model systems are bounded at least partly by solid walls. They can be fixed or moving (e.g., wall and stirrer of a stirred vessel). Walls considered here are impermeable for mass fluxes but may possibly conduct heat. To prevent convective fluxes through the wall the normal velocity component is set to zero or equal to the normal velocity of the moving wall. Assuming no slip conditions, the tangential velocity is equal to the wall velocity, which is zero with fixed walls. If slip is considered, the gradient of the tangential velocity components normal to the wall is zero. To avoid diffusive fluxes through the wall the concentration gradient normal to the wall is set to zero.
  For heat transfer different types of walls must be considered. The definition of temperature boundary conditions at walls is an approximation in most cases, because neither the physical conditions are well known there nor are they easy to control. Adiabatic walls are modeled by using a zero temperature gradient normal to the wall. At walls of fixed temperature this temperature is given as boundary condition. Heat fluxes across a wall can be described by using Neumann or Cauchy conditions.
- Outlet: No assumptions about the quantities of the transport values can be made for outlet conditions. However, it is assumed that the flow is uniform and no further sources occur. In that case velocities tangential to the outlet are zero. For all other unknowns (normal velocity, concentrations, temperature) zero gradients normal to the outlet are used. Furthermore, for incompressible flows the mass flow at each outlet must be specified to preserve the total mass.
  It is not always easy to find a location at which outlet boundary conditions can be specified with sufficient accuracy. Special problems occur in turbulent flows if vertex structures are not averaged but resolved and the assumptions for the velocity field do not hold. In such a case, boundary conditions for the pressure are stated instead of velocity conditions. The strong coupling between pressure and velocity described in context of the equation of motion allows one to solve these equations with this type of boundary conditions.
- Symmetry: Some systems show symmetries which allow one to reduce the model space and consequently the numerical effort. Symmetry can be utilized if it holds not only for the geometry, but also for the flow and for all transport variables.
  Symmetry is given when all gradients normal to the symmetry plane are zero and the velocity component normal to the symmetry plane is zero, too.

3.3. Turbulent Flow

While laminar flow is characterized by parallel stream lines, in a turbulent flow vortex structures of a large size range cause highly frequent fluctuations in all transport variables. All the equations introduced up to now are valid in laminar as well as in turbulent flows, as long as the conditions mentioned hold. The greatest problem in sim-

ulations of turbulent flows is that the resolution of the high-frequency fluctuations requires a fine numerical grid and small time steps. The number of grid cells N needed for a three-dimensional simulation can be estimated by

$$N=5^3 Re^{9/4} \tag{10}$$

where Re is the Reynolds number of the flow. Such simulations of turbulent flows, called *direct numerical simulations* (DNS) are possible only for small academic cases and relatively low Reynolds numbers with current (2004) computing power.

In most cases the exact structure of the turbulent flow is of minor interest and information on local averaged values is sufficient. Unfortunately, such an averaging can not be done in a simple way because the course-scale structure of the flow interacts strongly with the small-scale structure. Special attention must be given to chemically reactive flows. Chemical reactions can take place only if the species are mixed on the molecular scale. A model averaging this scale may overestimate the rate of fast chemical reactions.

Two general modeling approaches for turbulent flows are used nowadays: *Reynolds averaging* (Reynolds-averaged Navier–Stokes RANS) and *large-eddy simulations* (LES). RANS modeling is based on averaging all transport quantities over all scales of turbulent fluctuations. This allows one to use relatively coarse grids and can be applied to geometrically large systems with reasonable effort. The influence of turbulent structures is included in additional models. Large-eddy simulations work with a grid size which allows direct resolution of the large-scale part of the turbulence structures. Since the large eddies contain the majority of the turbulent energy, the return in accuracy is superproportional to the increase in effort with this approach. The filter width is a function of the grid spacing and can therefore be influenced by the user. Nevertheless, subgrid-scale models (SGS models) are required also for LES. The numerical effort for LES is significantly larger than for RANS simulations and can be handled at the moment only on high-performance parallel computers. Figure 2 gives an impression of the resolution of turbulent eddy structures by the different methods.

LES and RANS are illustrated in more detail for the equation of motion. Concentration and energy equations can be handled in a similar way, but all turbulence models are developed in the first step for momentum transport. Thus the mass and energy transport models use proportionality approaches, in most cases based on the parameters computed for momentum transport.

Reynolds Averaging. Reynolds averaging is based on the assumption that scales of the main flow and of turbulent fluctuations differ significantly. Therefore, each transport quantity can be split up into a time-averaged value $\overline{\phi}$ and a fluctuating value $\phi\prime$ in such a way that possible macroscale fluctuations of $\phi\prime$ are included in $\overline{\phi}$ while turbulent fluctuations are covered by $\phi\prime$:

$$\phi=\overline{\phi}+\phi' \tag{11}$$

In accordance with the goal of RANS to obtain averaged values, the transport equations are averaged. This yields for the Navier–Stokes equations:

$$\begin{aligned} \nabla\,\overline{\mathbf{v}} &= 0 \\ \rho\frac{\partial\overline{\mathbf{v}}}{\partial t} &= -\rho\overline{\mathbf{v}}\nabla\overline{\mathbf{v}}-\nabla\boldsymbol{\tau}_t+\mu\nabla^2\overline{\mathbf{v}}-\nabla\overline{p}+\rho\mathbf{g} \end{aligned} \tag{12}$$

where an additional term including the *Reynolds stress tensor* $\boldsymbol{\tau}_{\mathrm{t}}$ occurs.

$$\boldsymbol{\tau}_t=\rho\mathbf{v}'\times\mathbf{v}'=\begin{pmatrix} \rho\overline{v'_x v'_x} & \rho\overline{v'_x v'_y} & \rho\overline{v'_x v'_z} \\ \rho\overline{v'_y v'_x} & \rho\overline{v'_y v'_y} & \rho\overline{v'_y v'_z} \\ \rho\overline{v'_z v'_x} & \rho\overline{v'_z v'_y} & \rho\overline{v'_z v_z} \end{pmatrix} \tag{13}$$

With this term the averaged equation of motion is not closed because the velocity fluctuations are unknowns.

There are a large number of approaches to model the Reynolds stress tensor. To explain all of them is beyond the scope of this article. A detailed presentation is given in [2]. Therefore, only the most commonly used concept of the eddy viscosity approach together with the k—ε model will be sketched. The effect of the Reynolds stress tensor can be interpreted in analogy to the Newton stress tensor [3]. Therefore, it can be expressed proportionally to the gradient of the averaged velocity with a proportionality factor μ_{t} called *eddy viscosity*:

$$\boldsymbol{\tau}_t=\mu_t[\nabla\overline{\mathbf{v}}+(\nabla\overline{\mathbf{v}})^{\mathbf{T}}]-\frac{2}{3}\rho k\delta \tag{14}$$

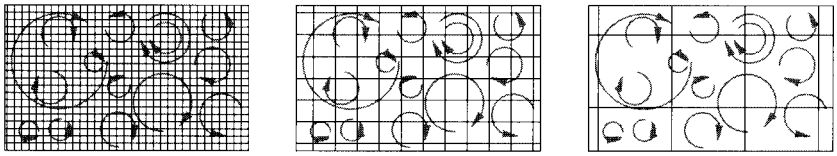

a) Direct numerical simulation b) Large eddy simulation c) Reynolds averaged simulation

Figure 2. Resolution of turbulent eddy structures by different modeling approaches
a) Direct numerical simulation; b) Large-eddy simulation; c) Reynolds-averaged simulation

where δ is the unit tensor and k is the turbulent kinetic energy defined by

$$k=\frac{1}{2}\overline{\mathbf{v}'\cdot\mathbf{v}'} \tag{15}$$

In contrast to the molecular viscosity, μ_t is the *eddy viscosity*, which is not a material constant but depends on the flow structure. The k—ε model expresses this dependence as a function of the turbulent kinetic energy k and its dissipation rate ε:

$$\mu_t=C_\mu\rho\frac{k^2}{\varepsilon} \tag{16}$$

with

$$\varepsilon=\frac{\mu}{2\rho}\overline{|\nabla\mathbf{v}'+\nabla\mathbf{v}'^{\mathbf{T}}|^2} \tag{17}$$

and k and ε computed by their own transport equations. The original model [4] uses the following form:

$$\begin{aligned}\frac{\partial\rho k}{\partial t} &= -\nabla(\overline{v}\rho k) + \nabla(\tfrac{\mu_t}{\sigma_k}\nabla k) + \hat{P} - \rho\varepsilon\\ \frac{\partial\rho\varepsilon}{\partial t} &= -\nabla(\overline{v}\rho\varepsilon) + \nabla(\tfrac{\mu_t}{\sigma_\varepsilon}\nabla\varepsilon) + C_{\varepsilon1}\tfrac{\varepsilon}{k}\hat{P} - C_{\varepsilon2}\rho\tfrac{\varepsilon^2}{k}\end{aligned} \tag{18}$$

with the production term $\hat{P}$ defined as

$$\hat{P}=-\frac{\mu_t}{2}|\nabla\overline{\mathbf{v}}+\nabla\overline{\mathbf{v}^{\mathbf{T}}}|^2 \tag{19}$$

The constants of the model are $C_\mu = 0.09$, $C_{\varepsilon1} = 1.44$, $C_{\varepsilon2} = 4.92$, $\sigma_k = 4.0$, $\sigma_\varepsilon = 1.3$.

The k—ε model is valid for fully developed isotropic turbulence. If this condition is not given, either extended RANS models or LES are required for an acceptable flow prediction.

Large-Eddy Simulations. A high-accuracy alternative to Reynolds averaging is given by large-eddy simulations. In this approach the transport values are filtered with a filter function $G(\mathbf{x},\mathbf{x}')$:

$$\hat{\phi}(\mathbf{x})=\int G(\mathbf{x},\mathbf{x}')\phi(\mathbf{x}')\,\mathrm{d}\mathbf{x}' \tag{20}$$

where G describes the influence of the value of ϕ in position $\mathbf{x}'$ on the filtered value $\hat{\phi}$ at position $\mathbf{x}$. Several forms of G are possible, but it is always a function of the grid size.

Filtering the equation of motion gives:

$$\rho\frac{\partial\hat{\mathbf{v}}}{\partial t}=-\rho\hat{\mathbf{v}}\nabla\hat{\mathbf{v}}-\nabla\boldsymbol{\tau}^s+\mu\nabla^2\hat{\mathbf{v}}-\nabla\hat{p}+\rho\mathbf{g} \tag{21}$$

In this equation the subgrid-scale Reynolds stress tensor $\boldsymbol{\tau}^s$ occurs. Its meaning is analogous to the Reynolds stress tensor in RANS but it considers only the effect of unresolved subgrid-scale structures.

Equation 21 can be solved only if $\boldsymbol{\tau}^s$ is expressed as a function of filtered values. For this purpose so-called subgrid-scale models (SGS models) are used.

SGS models of the Smagorinsky type are widespread [5]. They are similar to the eddy viscosity models in RANS, apart from the fact that they consider only sub-grid scale effects. In their structure they do not use information about the eddy structure of the resolved scales. Similarity models like that in [6] are based on the assumption that there is a similarity in the eddy structures of different scales. Therefore, the analysis of the resolved structure can be used to model the subgrid-scale. These models fail to describe dissipation at the smallest eddy scales, because this effect is not found at larger scales. So they have to be combined with other models to cover this aspect. Dynamic models, e.g., [7, 8] are based on this idea and extend upon it. For this approach two simulations on different grids are carried out. The SGS model is adapted to the differences in the predictions of both models. In this way parameters of the SGS scale model can be varied in time and space and dissipation effects are described. Such models are interesting

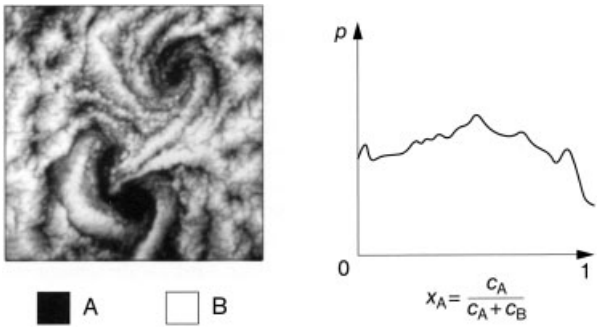

Figure 3. Modeling micromixing in a computational cell with probability density function approach
a) Distribution of species A (black) and B (white) in a turbulent flow with incomplete micromixing; b) Probability density function of mixing between A and B

for cases in which constant parameters are only a rough approximation. The major drawback of this method is its numerical effort, which is significantly higher than for the other methods.

Micromixing Models. If systems with fast chemical reactions are to be simulated, a description of the mixing state on the microscopic level is required. Neither Reynolds averaging nor large-eddy simulation provides such an description because they average or filter small-scale fluctuations. A remedy is the use of micromixing models which provide information about the micromixing quality as a function of turbulence quantities, like eddy viscosity or turbulent kinetic energy dissipation.

A widespread micromixing model is the probability density function approach (PDF) [9, 10]. It considers the possible mixing states of a fluid of certain composition and gives for each computational cell a probability density function for each state in such a away that the integral composition in this cell is satisfied. For a mixture of only two species A and B this is shown in Figure 3. The local probability density functions are computed from transport equations for the PDF.

Each mixing state is related to a certain chemical reaction rate. The integral of the reaction rates over the whole composition space gives the total reaction rate in the appropriate numerical cell.

Alternative micromixing models are

- Eddy break up model [11]: This is a simple empirical model which assumes binary chemical reaction with only one product. It is mainly applied for the modeling of combustion.
- Flamelet model [12, 13]: This model assumes chemical reaction in a two-dimensional laminar layer. It was developed for premixed and nonpremixed flames with a different definition of the reaction layer for the two cases.
- Engulfment model [14]: This model assumes that a fluid is mixed into another of different composition by formation of a rotational cylinder. Within the cylinder diffusion is the dominant mixing process and the composition changes proportionally to the change in cylinder size.

3.4. Multiphase Approaches

Multiphase systems are characterized by interfaces at which the fluid properties change discontinuously. This requires extensions of the model approaches introduced up to now for their application to multiphase systems. At the beginning of this chapter multiphase systems will be classified to describe the appropriate model for each type.

One type of classification refers to the geometric structure of the system (see Figure 4). If one phase consists of solid or fluid particles with a small size compared to the total system the system is called a *disperse system*. The particles form the disperse phase while the other phase is called the continuous one. For modeling disperse systems, an extended form of the continuum hypothesis is applied. Properties of both phases are considered to be continuously

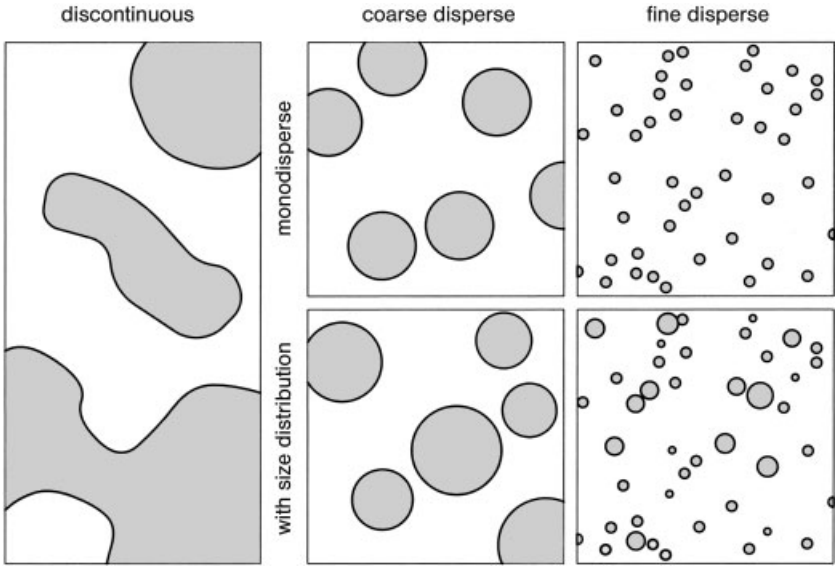

Figure 4. Geometric structure of multiphase systems

distributed over the whole system; specific properties of single particles are not relevant. The central property in that case is the volume fraction α which for phase *i* is defined as the ratio of the volume V_i of this phase and the total volume V:

$$\alpha_i = \frac{V_i}{V} \tag{22}$$

The sum of all volume fractions must be unity:

$$\sum_i \alpha_i = 1 \tag{23}$$

The assumption that the continuum hypotheses holds for disperse systems implies that changes of properties in a particle are not resolved. This means, e.g., in a temperature distribution that each particle is characterized by a mean temperature, but the different mean temperatures of all particles determine a particle temperature field of the system. If this approach is insufficient for the description of a system, e.g., because the temperature gradient at the interface is needed to determine the heat transfer between the phases exactly, it cannot be modeled with the approach for disperse systems.

An important property of disperse systems is the size distribution. Many phase interactions like buoyancy force or heat transfer rate depend on particle size, and therefore the particle size distribution is needed to determine these interactions exactly. The easiest case from the point of modeling are monodisperse systems; which all particles have the same size. For some applications the differences in the interactions do not need to be resolved, and this approach can be used. If the size distribution has to be considered the particle size becomes an additional independent variable in the system, which increases the complexity of the mathematical treatment significantly. Size distributions are relevant for all cases where the size range of the particles is very wide or the size distribution changes in the course of the process, e.g., by nucleation, evaporation, agglomeration, or similar processes. The computation of particle size distributions is based on the solution of population balances. This model approach is not discussed here further, but details can be found, e.g., in [15 – 17].

Another characteristic of a disperse system is the relative velocity between the phases. If the differences in gravitational forces are small compared to the momentum transfer between the phases, both phases have nearly the same velocity. This is the case for phases with similar densities or for systems with very small particles.

Another description refers to the interactions during momentum transfer. If particles move with the continuous phase but do not significantly influence the flow field of this phase it is called a *one-way interaction*. Systems with one-

way interactions do not have to be modeled with multiphase approaches. The flow field is modeled with the single-phase equations and the volume fraction can be handled like a concentration. The term *two-way interaction* is used, if there is a relative velocity between the two phases which causes a reciprocal momentum transfer. If, in addition, the volume fraction of the disperse phase is so high that interactions between the particles and also between particles and walls have to be considered, one speaks of *four-way* interactions.

If in a multiphase system a disperse phase cannot be defined clearly because the extension of all phases is too large to neglect gradients of the transport values in them, the system is called a *discontinuous multiphase system*. The way to model such a system is to use the approaches for single-phase systems and additionally to describe the condition at the interface. This works well if the extent of each single-phase region covers a significant part of the total domain.

The conditions at the interface act as boundary conditions for the balances in both phases. Therefore, two conditions must be defined for each transport variable in each point. One is the equilibrium of fluxes, and the other is the relation between the values of the transport quantity. As neither momentum nor mass or energy can be accumulated at the interface, the fluxes must be equal on both sides. The ratio of the values depends on the quantity considered. As a result of the nonslip condition the tangential velocities are equal. Unless there is a significant mass transfer over the interface (e.g., melting or evaporation) this holds also for the normal velocity; otherwise source terms must be considered in the stress balance at the interface. Thermal equilibrium at the interface states that the interfacial temperature for both phases is the same. The distribution of chemical species is not equal in equilibrium but described by a distribution coefficient. Detailed explanations and equations for such systems can be found, e.g., in [18, 19].

Euler–Euler Method. The Euler–Euler method is used for modeling disperse systems. It utilizes continuous transport equations for each phase. These equations are based on single-phase equations but weighted with the volume fraction and appended with source terms describing the phase interaction [20].

For the equation of motion the interactions between phases are forces. In all systems with a relative velocity between the phases the drag force plays a major role. It is the main force which prevents a continuous increase in the relative velocity between the phases. Further forces which have to be considered depending on the properties of the phases are the virtual mass force, the Basset force, and other forces caused by different relative velocities at different points of the particle surface, such as the Saffman force and the Magnus force.

Mass and heat transfer between the phases determine the interaction terms in the concentration and temperature equations. For systems with a significant mass transfer over the interface it must be considered that heat and momentum are also transferred together with that mass.

The Euler–Euler approach is preferably used for systems with high volume fractions of the disperse phase or for relatively large particle sizes, because there are no formal restrictions for these properties. Models for the changes in turbulence caused by phase interaction are not well developed yet. Size distributions can be described only with a high numerical effort because most forces change as a function of the particle size, and therefore the size distribution must be divided into size classes, whereby each class is handled as an individual phase.

Euler–Lagrange Method. Alternatively, the Euler–Lagrange approach can be used for disperse systems. In this approach the continuous transport equations are also used for the continuous phase and extended by interaction terms. The disperse phase is modeled differently: particle parcels are defined and distributed in such a way that they represent the particle phase with its volume fraction and properties. For each of these parcels *Pi* the path and the velocities (convective and rotational) are determined by ordinary differential equations:

$$\begin{aligned} \frac{d\mathbf{x}_{Pi}}{dt} &= \mathbf{v}_{Pi} \\ m_{Pi}\frac{d\mathbf{v}_{Pi}}{dt} &= \sum \mathbf{F}_i \\ I_{Pi}\frac{d\boldsymbol{\omega}_{Pi}}{dt} &= \mathbf{T}_i \end{aligned} \tag{24}$$

where $\mathbf{x}_{Pi}$ is the position of the parcel, $\mathbf{v}_{Pi}$ and $\boldsymbol{\omega}_{Pi}$ are its convective and rotational velocity,

$\mathbf{F}_i$ the forces acting between the faces, I_{Pi} is the moment of inertia, and $\mathbf{T}_i$ the torque.

The velocity of the continuous phase and the particle phase are corrected in an iterative procedure considering interactions based on current particle paths. The particle parcels only give a statistical representation of the particle phase if their characteristics are averaged over a reasonable time interval. The time interval for averaging has no limits in the stationary case aside from giving statistically reliable results. For transient investigations the time interval of averaging determines the time resolution of the simulation. If statistical reliability is insufficient, the time steps of the particle simulation must be reduced.

The forces for phase interaction in the second of Equations 24 are the same as those mentioned for the Euler–Euler approach. Mass and energy transfer also have to be considered along the particle paths and statistically evaluated.

Consideration of the particle paths reduces the volume fraction and particle sizes which can be handled by this method. The volume of a parcel presents can not be larger than the smallest grid cell. For high volume fractions the danger of having more than one parcel in a cell at a time further restricts the volume ratio between a cell and a parcel. Reasonable results are therefore only obtained for very low volume fractions, usually in the range of 1%, exceptionally up to 5%. The numerical effort of the method increases with the volume fraction, because additional parcels have to be introduced.

On the other hand, this method provides better models of the influence of particles on turbulence than the Euler–Euler approach. Defining parcels with different properties allows one to model size distributions with only a small additional effort. A detailed description of the Euler–Lagrange method is given in [21, 22].

Grid Adaptation to Interface. The best way to model a discontinuous multiphase system is to define the reference system in such a way that the interface has a fixed position in space. In that case the grid can be constructed so that the interface is located at grid lines. The interfacial conditions can then be applied easily.

In most cases interfaces move, and the reference system cannot be defined as described above. In this case the location of the interface is a part of the solution of the model. To incorporate the interfacial conditions as with boundary conditions the grid can be adapted to the new location of the interface. As long as the dislocation is small, this affects only the cell layer close to the interface and the numerical error introduced is small. For large dislocations the cells near the interface would be deformed severely and it is necessary to add or delete cells. Detailed descriptions of adaptive methods are given in [23].

Volume of Fluid. If there is a strong movement of the interface, the grid adaptation method is numerically laborious and erroneous. An alternative is to locate the interface at an arbitrary position of the grid. Different methods act in this way (segment method [24], marker and cell method [25]). The most popular method is the volume of fluid method [26]. A volume fraction is defined and a transport equation is solved for it similarly to the Euler–Euler method. The volume fraction in most cells is either zero or one and the interface is located in the cells where the volume fraction lies between these values. The exact location of the interface is determined by the volume fraction in the cell. The direction of the interface is normal to the gradient of the volume fraction.

4. Numerics

4.1. Basics

In chapter 3 mathematical models for the description of transport processes in fluids have been introduced. Clearly, the coupled system of integrodifferential equations and algebraic equations can be solved analytically only for a very small number of limiting cases. For all practically relevant situations only a numerical, computer-based solution is possible.

The basic idea of a numerical solution is to replace the partial differential equation which is continuously defined over the range in space and time with a system of algebraic equations which gives the solution only at defined discrete points or for discrete intervals. For the finite difference method (FDM), values at single points are computed; for the finite volume method (FVM), characteristic values are determined, which are constant for a numerical cell; and for the finite

element method (FEM), parameters of a polynomial function over a cell yield the solution. All these are available for separate times. A completely different approach, used with the lattice Boltzmann method, is described separately.

As the solution of the algebraic equations, called the discretized equations, is fixed to certain points or elements in space and time, the discretization of space and time is the first step to be performed. Time is a one-dimensional coordinate, and so it can simply be split into intervals, i.e., time steps. The time step size can vary over the total time period considered. Stability and accuracy of the procedure on the one hand and computing time on the other depend on the time-step size chosen in relation to the space discretization (see Section 4.2).

The discretization of the three-dimensional space offers much more possibilities. The two discretization methods mainly used in CFD, FVM and FEM, require the subdivision of the total model space into non overlapping cells. In the ideal case the grids formed from these cells meet the following requirements:

- Geometrical requirements: The outer border of the grid should fit as closely as possible to the physical borders of the system to introduce the boundary conditions at the correct location. In the case of simple geometrical structures this is not a problem but it might become one for jagged structures. The more variable the cell shapes a CFD code accepts, the more easily such structures can be described without extensive refinement close to the boundaries.
- Physical requirements: To achieve sufficient accuracy in the solution the grid size should increase in areas of steep gradients and especially where they are varying. This criterion has the major problem that in can be fulfilled precisely only if the solution is known. In many cases general physical understanding of the flow investigated is sufficient to estimate where local grid refinement is necessary. In other situations preliminary computations on a rough grid give an orientation for grid improvements.
  The simulation of turbulent flows requires a few grid cell layers parallel to the boundary for an appropriate modeling of the laminar sublayer.
- Numerical requirements: The numerical requirements depend mainly on the numerical method used and the way it is implemented in the code. The crux of the matter is whether a structured grid is needed or an unstructured one is allowed. Especially for unstructured grids a wide variety of cell shapes can be imagined but not every one is supported by every code. Nearly all unstructured solvers can handle tetrahedral and hexahedral cells separately or in combination in one grid. In some cases more general formulations of polyhedral cells are available but are not standard.
  From a numerical point of view serious jumps in the size of adjacent grid cells should be avoided. Furthermore, cells should not tend to degenerate, that is, the ratio of volume to surface should not become too small.

In many cases very large numbers of cells would be needed to meet all requirements sufficiently. This might push the computational effort to an unrealistic level. Therefore, a reasonable compromise has to be found.

The major distinction of grid types is made between structured and unstructured ones. A structured grid is defined by three bands of grid faces, where the faces of one band do not intersect with each other but with the faces from the other bands. There is no restriction on the shape of the faces or the coordinate system in which they are defined. The cell shapes in structured grids are hexahedral, sometimes with one collapsed cell face. Structured grids have only restricted applicability to complex geometries but for simple geometries they can be generated easily.

In unstructured grids no global grid faces or lines are defined, but the domain may be divided in any way. This offers many more possibilities for grid fitting, but it normally cannot be done without an automated grid generator.

For some simulations it is necessary to perform grid adaptation in the course of the simulation. Reasons might be, e.g., moving boundaries (as in stirred vessels), movements of interfaces which have to be tracked, or strong transient changes in the solution which require local changes in refinement. If the new structure of the grid depends on the numerical solution, the numerical effort for the adaptation strategies is relatively high. Nevertheless, in many cases

with grid adaptation a significant improvement in accuracy and efficiency of a simulation can be gained.

Discretization of the model equations can be based on the discretization of space and time. The different methods that are available vary in their mathematical approach and range of application. An insight is given here for two methods: the finite volume method, and the lattice Boltzmann method, with stress on the forme.

A third widespread method is the finite element method. It is extensively explained in → Mathematics in Chemical Engineering, Chap. 7.6 and therefore not discussed here.

4.2. Finite Volume Method

Historically, the most common numerical method in CFD is the finite volume method (FVM). It is also used by the presently most widest distributed commercial CFD codes FLUENT, CFX, and Star-CD. Its success in CFD is related to physical equivalents in the approach.

The fundamental concept is the consideration of grid cells (also called *control volumes*) in which all physical parameters and system characteristics are assumed to be constant. The balance equations are integrated over each of this cells. For volume sources these integrals can be solved directly by considering these sources as constant over the cell. The volume integrals of the transport terms are transformed into surface integrals by using the theorem of Gauss. As the physical properties are constant within each cell, they can also considered to be constant along each cell face, which permits the surface integral to be resolved. The value or the derivative of the solution at the cell face is contained in the expressions gained. While the values are constant within each cell, they jump unsteadily at the cell faces, and the values and derivatives at the faces have to be expressed by the values in the cells, which are fixed at certain reference points, i.e., the nodes.

This approach is, in general terms, the reverse of the derivation of the transport equations, where a finite model volume is considered initially and fluxes over the boundaries and volume sources are balanced. In the second step the limiting expression for the model volume approaching zero is determined. This analogy reveals another advantage of the method in a physical sense: as the surface integral of the transport terms expresses the fluxes over the cell surface, a numerical error in the determination of these fluxes does not violate the integral conservation of the balanced property. The laws of conservation of mass, momentum, and energy are the fundamentals of modeling for CFD, and therefore a method upholding these laws inherently has been most widely accepted by engineers.

Nowadays, this advantage of FVM is no longer significant, firstly, because the accuracy of all methods has been increased, so that for well-designed grids the error in the computation of fluxes is small compared to other errors in the course of simulations, and, secondly, because advanced grid techniques, like certain types of unstructured grids or sliding grids, do not uphold this advantage exactly. Nevertheless, it is still the most widely distributed CFD method and is therefore described here as an example. More details can be found, e.g., in [27 – 29].

Discretization in Space. We first consider the general shape of a transport equation for the quantity ϕ. For purposes of simplicity only the two-dimensional version is used, but the extension to three dimensions is straightforward. Discretization in time as discussed later, so we consider here the stationary form. The velocity field is assumed to be known as well as the molecular transport coefficient Γ, which might vary in a known way over the analyzed space.

$$0 = -v_x \frac{\partial \phi}{\partial x} - v_y \frac{\partial \phi}{\partial y} + \frac{\partial}{\partial x}\left(\Gamma \frac{\partial \phi}{\partial x}\right) + \frac{\partial}{\partial y}\left(\Gamma \frac{\partial \phi}{\partial y}\right) + S_\phi \tag{25}$$

For reasons of simplicity the integration is shown here for an equidistantly structured grid. But there is no general restriction — finite volume discretization can also be carried out for nonequidistant and unstructured grids. Note that in the two-dimensional case the "volume" for integration is a face and the "volume surface" is a line.

For the grid the compass notation shown in Figure 5 is normally used [27]. The nodes are marked with capital letters, *P* being that of the cell considered. The points at cell faces are marked by lower-case letters. The final linear equation system can contain only values with

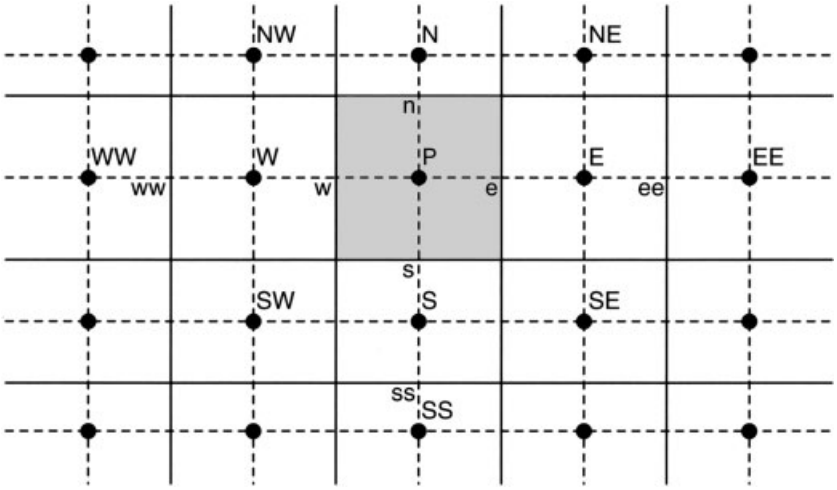

Figure 5. Nomenclature in a two-dimensional grid for finite volume method

capital letter indices, values with lower case indices are not defined and must be replaced by the former.

Integration of Equation 25 over the control volume gives

$$0 = -\int_V v_x \frac{\partial \phi}{\partial x} dV - \int_V v_y \frac{\partial \phi}{\partial y} dV + \int_V \frac{\partial}{\partial x}\left(\Gamma \frac{\partial \phi}{\partial x}\right) dV + \int_V \frac{\partial}{\partial y}\left(\Gamma \frac{\partial \phi}{\partial y}\right) dV + \int_V S_\phi dV \tag{26}$$

Conversion of volume integrals of the transport terms and resolution lead to:

$$0 = -\int_A \mathbf{n} v_x \phi dA - \int_A \mathbf{n} v_y \phi dA + \int_A \mathbf{n}\left(\Gamma \frac{\partial \phi}{\partial x}\right) dA + \int_A \mathbf{n}\left(\Gamma \frac{\partial \phi}{\partial y}\right) dA + \int_V S_\phi dV$$

$$0 = -(v_x A\phi)_e + (v_x A\phi)_w - (v_y A\phi)_n + (v_y A\phi)_s + \left(\Gamma A \frac{\partial \phi}{\partial x}\right)_e - \left(\Gamma A \frac{\partial \phi}{\partial x}\right)_w + \left(\Gamma A \frac{\partial \phi}{\partial y}\right)_n - \left(\Gamma A \frac{\partial \phi}{\partial y}\right)_s \overline{S} V \tag{27}$$

A_e, A_w, A_n and A_s are the cell faces crossing e, w, n and s and **n** is the face normal.

Finally, the values and the derivatives of ϕ at the cell faces must be replaced as a function of the values at the nodes. The derivatives are approximated by using the central difference, which is accurate to the second order, e.g.

$$\left(\Gamma A \frac{\partial \phi}{\partial x}\right)_w = \Gamma_w A_w \left(\frac{\phi_P - \phi_W}{\delta x_{WP}}\right) \tag{28}$$

where δx_wp is the distance between the points W and P.

To express the values of ϕ various approaches are used which differ in their stability and accuracy. Most of them are asymmetric and depend therefore on the flow direction. We discuss them for cell face w and a velocity $\nu_{\mathrm{x,w}}$ from W to P.

- *Upwind Differencing Scheme (UD)*

$$\phi_w = \phi_W \tag{29}$$

This scheme is unconditionally stable, but it is only accurate for the first order. It is the only scheme which does not produce any overshots or undershots without additional limiters and therefore guarantees the solution to be in the physically meaningful range (no negative concentrations, etc.). On the other hand, the numerical error is relatively large. It has the effect of smoothing strong changes in the gradients—a property called *numerical diffusion*. The error can be minimized if the local velocity vector is nearly parallel to the face normal vector.

- *Central Differencing Scheme (CD)*

$$\phi_w = \frac{\delta x_{Ww}\phi_W + \delta x_{wP}\phi_P}{\delta x_{WP}} \tag{30}$$

This scheme is only conditionally stable but it is second-order. Central schemes normally give a better approximation of steep slopes, but they show significant nonphysical oscillations of the solution. In the worst case these oscillations can build up and lead to a diverging solution. CD schemes are stable and advisable for diffusion-dominated processes.

- *Quadratic Upstream Interpolation for Convective Kinetics (QUICK)*

$$\phi_w = -\frac{1}{8}\phi_{WW} + \frac{6}{8}\phi_W + \frac{3}{8}\phi_P \tag{31}$$

QUICK is the basis for most third-order schemes used today. With a combination of central and upwind aspects it joints properties of both schemes discuss above. Oscillations occur, but to a lesser extent than with the CD scheme. QUICK has a much wider range of stability than CD but restrictions remain for systems without significant diffusion terms. To improve this third-order scheme various advancements have been developed. The main advantage is gained from the combination with so-called *limiters*. Limiters are extensions of the numerical scheme which prevent the formation and pronunciation of local extrema but keep the numerical accuracy of a scheme.

For the source term a linear approach in ϕ is used:

$$\overline{S}V = S_0 + S_1\phi \tag{32}$$

Discretization in Time.

For transient cases a time derivative of the transport quantity also occurs in the equation. If the two-dimensional Equation 25 is extended in this way one obtains:

$$\frac{\partial\phi}{\partial t} = -v_x\frac{\partial\phi}{\partial x} - v_y\frac{\partial\phi}{\partial y} + \frac{\partial}{\partial x}\left(\Gamma\frac{\partial\phi}{\partial x}\right) + \frac{\partial}{\partial y}\left(\Gamma\frac{\partial\phi}{\partial y}\right) + S_\phi \tag{33}$$

The philosophy for the time discretization is the same as for the discretization in space: The equation is integrated over the grid cell as well as over the time interval. The integration of the left-hand side can be performed easily:

$$\int_V\int_t^{t+\Delta t}\frac{\partial\phi}{\partial t}dtdV = V(\phi_P^{t+\Delta t} - \phi_P^t) \tag{34}$$

For the time integral of the right-hand side a linear relation to the values of ϕ at the old and the new time step is assumed:

$$\int_t^{t+\Delta t}\phi dt = \left[f\phi^{t+\Delta t} + (1-f)\phi^t\right]\Delta t \tag{35}$$

where f is a weighting factor between zero and one. For a space discretization which needs only the values of ϕ in P and its direct neighbor points the discretized equation reads:

$$\begin{aligned} a_P^{t+\Delta t}\phi_P^{t+\Delta t} = {} & a_E\left[f\phi_E^{t+\Delta t} + (1-f)\phi_E^t\right] + \\ & + a_W\left[f\phi_W^{t+\Delta t} + (1-f)\phi_W^t\right] \\ & + a_N\left[f\phi_N^{t+\Delta t} + (1-f)\phi_N^t\right] + \\ & + a_S\left[f\phi_S^{t+\Delta t} + (1-f)\phi_S^t\right] + \\ & + a_P^t\phi_P^t \end{aligned} \tag{36}$$

in which the coefficients $a_{\mathrm{P}}^{t+\Delta t}$ and a_{P}^t depend on f linearly, and a_{i} are the coefficients resulting from space discretization and from Equation 34.

For different values of f different methods can be derived. The most common ones are shown in Figure 6 .

- *f=0 (explicit method)*
This method is numerically simple to use because only independent linear equations are derived. It is of first-order accuracy. On the other hand, it is only conditionally stable. For applications in which the flux is dominated by the convective part (and this is the major part of all CFD applications) the *Courant–Friedrich–Lewy criterion* determines the stability range:

$$1 \geq C_\Delta = \frac{|\mathbf{v}|\,\Delta t}{\Delta x} \tag{37}$$

C_Δ is the so-called *Courant number*.

- *f=1 (implicit method)*
This method is unconditionally stable and also first-order accurate. If a final state of a process is more interesting than the transient behavior, this method can be used with relatively large time steps.

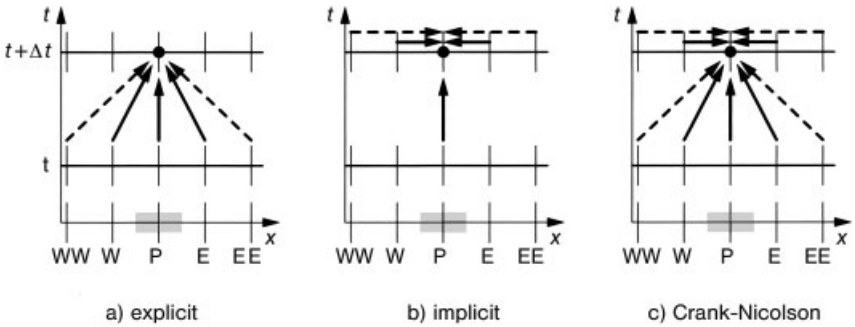

Figure 6. Discretization in time – coupling of values according to different methods
a) Explicit; b) Implicit; c) Crank-Nicolson

- *f=0.5 (Crank–Nicolson method)*
  This method is second-order accurate. It is conditionally stable but larger Courant numbers than in the explicit method can be handled. It is used if a high accuracy in time is required.

4.3. Pressure Correction Methods

One of the crucial points of fluid dynamic modeling are the nonlinearities in the equation of motion and the lack of an explicit pressure equation for incompressible flow. This was outlined in connection with the equation of motion.

There are different approaches to deal with this problem. The directly coupled solution of the equation of motion and the equation of continuity is possible, but is mathematically unstable, especially for the finite volume method, and is seldom used. Also, the approach of an artificial compressibility is only successful and efficient in a few cases.

The most commonly used alternatives are pressure correction schemes. They break up the velocity and pressure into an estimated value and a corrective part. Starting with an estimated pressure field, they compute an estimated velocity field, or vice versa, using the equation of motion. For the velocities in front of the derivatives in the convective terms (see Equation 6) estimates are used. Reorganization of the discretized momentum balance gives an equation for the pressure correction as a function of velocity correction. With the help of this equation the discretized continuity equation can be transformed into a pressure correction equation. Iterative solution of all these equations lead to a stepwise approach of velocity and pressure. It is important to update the "parameter velocity" in the convective term and other dependent parameters in each iterative cycle. If necessary, coupled equations (e.g., turbulence models) must be solved at the end of each step.

One of the most basic pressure correction algorithms is the SIMPLE algorithm [27]. Because it does not converge stably, a significant under-relaxation has to be used. Reversed algorithms have been developed since then, but the underlying basic principles are the same.

4.4. Lattice Boltzmann Method

The lattice Boltzmann method (LBM) differs greatly from FVM because it is not a discretization of the transport equations but a statistical method which gives an average result for the solution of the equations. To provide sufficient statistics the method requires a relatively fine grid and small time steps, but it performs much faster on a fine grid than FVM does on the same grid. The grid used is completely regularly structured, but the location of wall boundaries is not restricted to the grid lines. Therefore, the method is suited for geometries with sophisticated wall shapes like porous media. While some years ago the grid requirements caused an impassable numerical effort, such a method can be handled today on parallel computers. It is especially suited for simulations which require a fine grid to handle the models, like large-eddy simulations of turbulent flows.

For lattice Boltzmann simulations a number of numerical particles with properties like velocity, temperature, and concentration are dis-

tributed over the grid in such a way that averaging of their properties gives the initial conditions of the flow. The simulation consists of a first step in which the particles move due to their individual velocities. In a second step collisions between particles and also between particles and walls are interpreted according to collision rules. The collision rules describe the change in the properties of each of the particles in the collision. They must be formulated in such a way that they cause statistical changes so that the averaged property field fulfills the transport equations. [30, 31]

As the method is still relatively new, not all models can be used yet, but the current development is rapid.

5. Interpretation

As mentioned in Chapter 2, the volume of computed data of a CFD simulation is too large to interpret the numbers directly. In a graphic visualization the data are much easier to comprehend.

A basic step of a suitable visualization is an appropriate reduction of the data. Three-dimensional transient data cannot be visualized at once, and the visualization of the velocity vector field can not be realized effectively in a two-dimensional plane. But not all aspects of the data field are relevant to answering the initial question. Relevant data, however, can be extracted, and data can be averaged or combined to derive meaningful results. To transfer information into a graphical presentation requires that the figure be clearly readable and quantitative. Aspects of data reduction (location of a slice, chosen velocity components) have to be marked, and a legend must be given which allows quantitative interpretation of colored figures.

Visualization is not only useful for finding the result of the simulation. It also helps to detect errors. For this purpose the whole data field should at least be scanned roughly. As the aim of this procedure is to find irregularities in the solution, the layout of the presentation is of minor interest, but a quantitative interpretation of the data must be possible.

To judge the results all errors of the simulation must be analyzed as well as possible and quantified. The main categories are:

- *Modeling errors* refer to the differences between the values of all quantities in a real physical system and the values resulting from an exact solution of the mathematical model. Not all physical relations can be reproduced exactly in a mathematical formulation. Furthermore, not all mathematical models can be solved numerically with reasonable effort. Therefore, additional assumptions and restrictions are introduced which increase the modeling error. Examples are the application of turbulence models, the neglect of temperature dependencies of parameters or simplified approaches for boundary conditions.
  A quantitative estimate of the modeling errors is very difficult. If the mathematical model can be solved analytically (at least for a limiting case) and the corresponding quantities can be measured, a direct comparison is possible. Even in this case the measurement error from the experiment must be considered. In most cases such a direct comparison is not possible and one is restricted to experience from similar cases or to comparison of different model approaches for the same situation.
- *Discretization errors* refer to the difference between the solution of the continuous mathematical model and the discretized equations. The discrete solution is computed only for the nodes (which are representative for a certain volume according to FVM) and for time steps. For the locations between these points in space and time only approximations based on the values at the nodes can be used.
  The solution of the discrete equations at the nodes differs from the solution of the continuous equations at the same points. This is caused by the assumptions and approaches used in the course of discretization. While this is valid for all discretization schemes, it can be most easily explained for the finite difference method (FDM) (see → Mathematics in Chemical Engineering, Chap. 7.3 or [32]).
  According to this method, the approximation is derived from the Taylor series where only the first terms are considered. The neglected terms give the discretization error. The procedure is based on the assumption that the following terms in the Taylor series are decreasing. Therefore, the discretization error can be approximated by the first neglected term. This

term is proportional to the grid size with a certain power. This power is the so-called *order of accuracy*. The more terms considered the higher the order of the scheme. The order of a scheme is also used for other discretization methods. It states by which power the discretization error is reduced in relation to the reduction of the grid size. This means that a scheme of higher order is not always more accurate than a scheme of lower order if the schemes are derived in completely different manners. But the influence of grid size reduction is greater for the scheme of higher order.

- *Truncation of iterative procedures*: Iterative procedures are used on different levels for the solution of the discrete equations. Normally the linear algebraic equation systems are solved iteratively. Also, the coupling between the equations is often handled iteratively. The most important example of this are the pressure correction schemes (see Section 4.3). Iterative procedures approach the solution asymptotically. They are terminated if a certain accuracy is reached. Criterion for termination is the residual. It can be defined in two different ways. The first alternative is the difference in the solution between two iterations. This value is easily determined, but it is dangerous, because for slow convergence the differences between two iterations is small, even far away from the solution. The second alternative can be used if the problem can be formulated as:

$$\mathbf{F}(\mathbf{x}) = 0 \tag{38}$$

The residual of the nth iteration $\mathbf{r}^n$ is then defined as:

$$\mathbf{F}(\mathbf{x}^n) = \mathbf{r}^n \tag{39}$$

Both definitions of the residual give values for each node. For a truncation criterion different statistical methods are used to derive a characteristic value. This can be either the maximum, the average, or a weighted average of the value itself or of a normalized value.

- *Representation errors*: In the computer all numbers are represented by a limited number of digits. This might result in significant errors, e.g., if a small difference of large numbers is computed. Normally, this type of error is of minor importance in CFD simulations.

The different types of errors are usually additive. Therefore, it does not help to improve the accuracy of one aspect if the error in this regard is smaller than other errors. In most cases for simulation the limiting factor for error reduction is the modeling error. All other errors are usually small by comparison. A quantitative judgment has a higher priority than indifferent formulations such as "grid independence".

6. Industrial Application

CFD is a tool of increasing importance in industrial applications. The expertise of industrial users is mainly in the field of fluid dynamics or chemical engineering, not in numerics or programming. Therefore, industrial problems are solved with commercial codes which are available at high level with regard to modeling and numerics. Furthermore, such codes have a user-friendly interface and support which allows even inexperienced users to obtain first results quite fast. Commercial codes are continuously being developed and achieve more or less the state of the art. Nevertheless, the different codes are specialized in different fields of applications and therefore it is necessary to check the available models before a new kind of application is tackled. Often additional models can be implemented by user coding, but the effort required to do so is great and it should be checked before beginning whether another code provides a better foundation for the case considered.

To illustrate different modeling requirements, some examples for applications in chemical engineering are presented. This cannot give a complete overview of the most current usage of CFD in chemical engineering but gives an impression of the variety of applications.

- Tubular reactor: Tubes without internals have a simple geometric structure. They are a part of nearly all technical apparatuses as inlet and outlet pipes and in some cases they are used as reactors. Problems in modeling occur due to the large ratio between length and diameter, which requires a compromise in meshing between number of cells and shape of cells. In some cases the number of cells can be reduced by utilizing symmetry properties of the system.

Tubular reactors are often used to mix non-Newtonian fluids or, in a jet configuration, to mix Newtonian fluids very fast down to microscopic levels. For non-Newtonian fluids, e.g., in polymerization processes, chaotic mixers [33] consisting of tubular loops or static mixers [34, 35] within the tubes are used. In both cases the geometry of the flow becomes much more complicated. Especially for static mixers sufficient flexibility is only provided by unstructured grids.

In jet flows the modeling of turbulence is of great importance. As tubular reactors are relatively narrow, wall effects have a significant influence on the flow. Therefore, in most cases special wall models have to be used, and this increases the modeling and numerical effort. Furthermore, turbulence in tubes is often nonisotropic. This reduces the applicability of RANS simulations.

- Stirred vessel: Stirred vessels are widely used in chemical engineering for single phase and multiphase flows. The flow field is markedly inhomogeneous and transient. Characteristic of all stirred vessel configurations is the combination of static and moving elements. In the simplest case the static parts are rotationally symmetric. Then they can be modeled as a moving wall in a rotating reference frame [36]. If there are baffles, the rotational symmetry is broken and extended approaches are used. A good overview on these approaches is given by in [37]. If only the stationary solution averaged over a large number of stirrer revolutions is at interest, the power input from the stirrer can be included in source terms. This system can be solved for the stationary case on a fixed grid. For more detailed simulations the grid is divided in two overlapping or nonoverlapping parts. One of them moves with the stirrer; the other is static. As the stirrer motion has to be resolved sufficiently in time, only small time steps can be used, and the simulation becomes relatively extensive.

  RANS simulations often do not give a sufficient description, especially in presence of chemical reactions, because the flow field is inhomogeneous. As the fine resolution required for large-eddy simulations is numerically extensive, a remedy is to apply the lattice Boltzmann method [38].
- Precipitation reactor: Precipitation is usually simulated to predict the particle size distribution in the system [39, 40]. The initial supersaturation in precipitation processes is high, and therefore nucleation and particle growth start at a high rate. This requires appropriate turbulence modeling, including micromixing models [41]. The particles are very small, so only one-way coupling is relevant and the simulation can be carried out as pseudo-single phase. On the other hand, the solution of the population balance must be integrated into the simulation [15].
- Bubble column: Bubble columns are a typical example for two-phase flows. The difference in density between gas and liquid causes a relevant relative velocity between the phases. Depending on bubble size and volume fraction, either Euler–Euler [42, 43] or Euler–Lagrange [44, 45] approaches are used. In most practical applications the requirements for the Lagrangian approach are not met because the volume fraction is too high. But to describe the nonuniform bubble size population, balances have to be solved. This is much easier in connection with the Lagrangian approach than with the Eulerian approach.
- Membrane module: A newer apparatus used in chemical engineering is the membrane module. Its numerical simulation is at the very beginning because of the variety of aspects which have to be considered. These are, for instance:

  Presently the modeling approaches do not consider all of these aspects. The main stress is on modeling the two flow regions divided by the porous membrane under certain pressure conditions [46, 47].
  - Flow through a porous medium
  - Coupling of two flow regions
  - Multiphase flow
  - Non-Newtonian fluids
  - Fouling

It is obvious from this listing that the qualified application of CFD can not be performed without well-founded knowledge of applicable models and some insight into the numerical procedures.

The development of refined products like fine chemicals and active agents will be of growing importance in the chemical industry, and CFD

will play an important part in the design of the necessary apparatus. This requires the consideration of parallel and consecutive chemical reactions of different rates and simulations of multiphase systems including the prediction of particle size distributions. Consequently, emphasis in CFD development for this field will be on models and solution methods for fine spatial and temporal resolution and for a detailed description of the physics.

The advances in CFD interact strongly with hardware development. On the one hand the available computing power determines which mathematical models and which numerical methods can be used with reasonable numerical effort, and on the other it affects the direction of development of new models and methods. Due to the fast pace at which new powerful computer hardware evolves and becomes available, model development does not restrict itself to quantitative refinement but regularly includes new fields of research. More recent improvements are large-eddy simulations of turbulent flows instead of Reynolds-averaged approaches, and models for multiphase systems. A newer trend in the field of numerical methods is the application of the lattice Boltzmann method. Further innovations can be expected.

7. References

1. R.B. Bird, W.E. Steward, E.N. Lightfoot: *Transport Phenomena*, 2nd ed., John Wiley & Sons, New York 2002.
2. S. Pope: *Turbulent Flows*, Cambridge University Press, Cambridge 2000.
3. J.V. Boussinesq: "Essa sur la theories des eaux courantes", *Mém. prés. par div. savants à l'acad. sci. de Paris* **23** (1877) 1.
4. B.E. Launder, D.B. Spalding: "The Numerical Computation of Turbulent Flows", *Comput. Methods Appl. Mech. Eng.* **3** (1974) 269.
5. J. Smagorinsky: "General Circulation Experiments with the Primitive Equations, Part I: The Basic Experiment", *Monthly Weather Rev.* **91** (1963) 99.
6. J. Bardina, J.H. Ferziger, W.C. Reynolds: Improved Turbulence Models Based on Large Eddy Simulation of Homogeneous, Incompressible, Turbulent Flows, Technical Report TF-19, Thermal Sciences Div., Dept. of Mech. Eng., Stanford University, Stanford, CA. 1980.
7. M. Germano, U. Piomelli, P. Moin, W.H. Cabot: "A Dynamic Subgrid Scale Eddy Viscosity Model", in Proc. Summer Workshop, Center for Turbulent Research, Stanford CA 1990.
8. C. Meneveau, T.S. Lund, W.H. Cabot: "A Lagrangian Dynamic Subgrid-Scale Model of Turbulence", *J. Fluid Mech.* **319** (1996) 353.
9. J. Baldyga: "A Closure Model for Homogeneous Chemical Reactions", *Chem. Eng. Sci.* **49** (1994) 1985.
10. R.O. Fox: "On the Relationship between Lagrangian Micromixing Models and Computational Fluid Dynamics", *Chem. Eng. Process* **37** (1998) 521.
11. B.F. Magnussen, B.W. Hjertager: "On the Structure of Turbulence and a Generalised Eddy Dissipation Concept for Chemical Reaction in Turbulent Flow" in *19th AIAA Aerospace Meeting*, St. Louis, USA, 1981.
12. A. Linan: On the Internal Structure of Laminar Diffusion Flames, Technical Note, Inst. nac. de tec. aeron., Esteban Terradas, Madrid, Spain, 1961.
13. N. Peters: "A Spectral Closure for Premixed Turbulent Combustion in the Flamelet Regime", *J. Fluid Mech.* **242** (1992) 611.
14. J. Baldyga, J.R. Bourne: *Turbulent Mixing and Chemical Reactions*, John Wiley & Sons Ltd., Chichester 1999.
15. A.R. Paschedag: *CFD in der Verfahrenstechnik: Allgemeine Grundlagen und mehrphasige Anwendungen*, Wiley-VCH, Weinheim, Germany 2004.
16. D. Ramkrishna: *Population Balances. Theory and Application to Particulate Systems in Engineering*, Academic Press, San Diego 2000.
17. A. Gerstlauer, A. Mitrovic, S. Motz, E.-D. Gilles: "A Population Balance Model for Crystallization Processes using two Independent Particle Properties", *Chem. Eng. Sci.* **56** (2001) 2553.
18. W.M. Deen: *Analysis of Transport Phenomena*, Oxford University Press, New York 1998.
19. J.C. Slattery: *Advanced Transport Phenomena*, Cambridge University Press, New York 1999.
20. D.A. Drew: "Mathematical Modeling of Two-Phase Flow", *Ann. Rev. Fluid Mech.* **15** (1983) 261.
21. C.T. Crowe., M. Sommerfeld, Y. Tsuji: *Multiphase Flows with Droplets and Particles*, CRC Press, Boca Raton 1998.

22. M. Sommerfeld, W. Krebs: "Particle Dispersion in a Swirling Confined Jet Flow", *Part. Part. Syst. Characterization* **7** (1990) 16.
23. J.H. Ferziger, M. Peric: *Computational Methods for Fluid Dynamics*, 2nd rev. ed., Springer, Berlin 1999.
24. B.D. Nichols, C.W. Hirt: "Calculating Three-Dimensional Free Surface Flows in the Vicinity of Submerged and Exposed Structures", *J. Comput. Phys.* **8** (1971) 434.
25. F.H. Harlow, J.E. Welch: "Numerical Calculation of Time-Dependent Viscous Incompressible Flow of Fluid with Free Surface", *Phys. Fluids* **8** (1965) 2182.
26. C.W. Hirt, B.D. Nichols: "Volume of Fluid (VOF) Method for the Dynamics of Free Boundaries", *J. Comput. Phys.* **39** (1981) 201.
27. S.V. Patankar: *Numerical Heat Transfer and Fluid Flow*, Hemisphere Publishing Co., Washington 1980.
28. H.K. Versteeg, W. Malalasekera: *An Intr. to Computational Fluid Dynamics. The Finite Volume Method*, Longman, Harlow 1995.
29. C.A.J. Fletcher: *Computational Techniques for Fluid Dynamics*, vols. **1** and **2**, Springer, New York 1988.
30. U. Frisch, D. d'Humieres, B. Hasslacher: "Lattice Gas Hydrodynamics in Two and Three Dimensions", *Complex Syst.* **1** (1987) 649.
31. Succi S.: *The Lattice Boltzmann Equation for Fluid Dynamics and Beyond*, Oxford University Press, Oxford 2001.
32. J.D. Anderson Jr.: *Computational Fluid Dynamics: the Basics with Applications*, McGraw-Hill, New York 1998.
33. A. Birtigh, G. Lauschke, W.F. Schierholz, D. Beck, Ch. Maul, N. Gilbert, H.-G. Wagner, C.Y. Werninger: "CFD in der chemischen Verfahrenstechnik aus industrieller Sicht", *Chem. Ing. Tech.* **72/3** (2000) 175.
34. Th. Avalosse, M.J. Crochet: "Finite Element Simulation of Mixing: 2. Three-Dimensional Flow through a Kenics Mixer", *AIChE J.* **43/3** (1997) 563.
35. E.S. Mickaily-Huber, F. Bertrand, P. Tanguy, T. Meyer, A. Renken, F.S. Rys, M. Wehrli: "Numerical Simulations of Mixing in an SMRX Static Mixer", *Chem. Eng. J.* **63** (1996) 117.
36. F. Bertrand, P.A. Tanguy, E. Brito de la Fuente, P. Carreau: "Numerical Modelling of the Mixing Flow of Second-Order Fluids with Helical Ribbon Impellers", *Comput. Meth. Appl. Mech. Eng.* **180** (1999) 267.
37. A. Brucato, M. Ciofalo, F. Grisafi, G. Micale: "Numerical Prediction of Flow Fields in Baffled Stirred Vessels: A Comparison of Alternative Modelling Approaches", *Chem. Eng. Sci.* **53/21** (1998) 3653.
38. J. Derksen, H.E.A. Van den Akker: "Large Eddy Simulations on the Flow Driven by a Rushton Turbine", *AIChE J.* **45/2** (1999) 209.
39. D.L. Marchisio, A.A. Barresi, R.O. Fox: "Simulation of Turbulent Precipitation in a Semi-batch Taylor-Couette Reactor Using CFD", *AIChE J.* **47/3** (2001) 664.
40. A.R. Paschedag: "Modelling of Mixing and Precipitation Using CFD and Population Balances", *Chem. Eng. Tech.* **27/3** (2004) 232.
41. D.L. Marchisio: Precipitation in Turbulent Fluids, PhD thesis, Politecnico di Torino, 2002.
42. A. Sokolichin, G. Eigenberger: "Applicability of the Standard $k-\varepsilon$ Turbulence Model on the Dynamic Simulation of Bubble Columns: Part I Detailed Numerical Simulations", *Chem. Eng. Sci.* **54** (1999) 2273.
43. L.I. Zaichik, V.M. Alipchenkov: "A Kinetic Model for the Transport of Arbitrary Density Particles in Turbulent Shear Flows", in: Proc. Turbulence and Shear Flow Phenomena 1, Santa Barbara 1999.
44. A. Lapin, A. Lübbert: "Numerical Simulation of the Dynamics of Two-Phase Gas-Liquid Flows in Bubble Columns", *Chem. Eng. Sci.* **49** (1994) 3661.
45. L. Sanyal, S.V. Squez, S. Roy, M.P. Dudukovic: "Numerical Simulation of Gas-Liquid Dynamics in Cylindrical Bubble Column Reactors", *Chem. Eng. Sci.* **54** (1999) 5071.
46. D.E. Wiley, D.F. Fletcher: "Techniques for Computational Fluid Dynamics Modelling of Flow in Membrane Channels", *J. Membrane Sci.* **211** (2003) 127.
47. C.A. Serra, M.R. Wiesner: "A comparison of Rotating and Stationary Membrane Disk Filters using Computational Fluid Dynamics", *J. Membrane Sci.* **165** (2000) 19.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Molecular Dynamics Simulation

Standard Article

Philippe A. Bopp[1], Jörn B. Buhn[2], Manfred J. Hampe[3]
[1]Université Bordeaux I, France
[2]Technische Universität Darmstadt, Darmstadt, Germany
[3]Technische Universität Darmstadt, Darmstadt, Germany

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.i16_i01
Article Online Posting Date: August 15, 2004

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Fundamental Approximations
3. Interaction Models
3.1. Intramolecular Interactions
3.2. Intermolecular Interactions
3.2.1. Pair Potentials
3.2.2. Three-Body and Higher Order Intermolecular Potentials
3.3. Example: Intermolecular Potentials for an Aqueous Salt Solution
3.4. Data for Interaction Models
4. Approximations for the Simulations
5. Constructing the Ensemble
5.1. Method
5.2. Setting up a Simulation
6. Obtaining Results
6.1. Simple Averages
6.2. Time-Correlation Functions
7. Examples
8. Applications

[Top of Page]

1. Introduction

More than a century ago LUDWIG BOLTZMANN (1844–1906) elaborated a procedure allowing the macroscopic properties of matter to be linked with the molecular details and interactions of its constituent particles. This approach is known today as statistical thermodynamics or, more generally, as statistical mechanics. However, up to the arrival of sufficiently cheap and accessible numerical computers in the last third of the twentieth century, practical results could be obtained only for the simplest systems, in some limiting cases, or under drastic assumptions. This has proven intellectually challenging and has yielded many insights, but the urge has remained to overcome the simplifications and to look at specific systems of interest in various disciplines such as physics, chemistry, biology, geoscience, and engineering. The limitations have been pushed back in the last few decades, and many systems can now be investigated by following Boltzmann's principles. Nevertheless, it is still impossible to carry out this method for any real system of even the smallest macroscopic size. Approximations, although much weaker than before, must still be made, and the task remains daunting in many cases.

The computational methods used today to follow Boltzmann's lead, i.e., to cross the bridge between the microscopic, molecular world and the macroscopic observables measured in the laboratory, are known as molecular simulations. They come in many variants, the two major families of which are Monte-Carlo (MC) simulations [1] and the molecular dynamics (MD) simulations. The basic idea underlying all molecular simulations is simple and follows directly from Boltzmann's considerations: Construct a sufficient number of microscopic configurations, or states, compatible with (1) knowledge about the intermolecular (or interatomic) interactions in the system and (2) the macroscopic thermodynamic constraints (temperature, density, etc.) of the system under consideration. This could be called the “Gallup poll” [2] approach to the

thermodynamic ensemble: Since this powerful abstract notion upon which the Boltzmann approach rests cannot be practically realized, one must be satisfied with a representative sample. Like in the polls, the trouble lies, of course, in the representativeness. However, once the sample has been constructed, it can be used to do statistics, i.e. to compute average values of quantities that can be calculated from the configurations in these samples.

Example 1. Assume that the positions of all atoms in a sufficiently large sample of water are known over a sufficient length of time under the external constraints (e.g., at a given total energy, temperature, or . . .) imposed on the system. We want to know the thermodynamic average of all distances between neighboring pairs of oxygen atoms, d_{O-O}, under these conditions. Since the positions of the atoms have been constructed by means of the simulation so as to be compatible with the external conditions, and there are enough of them, according to Boltzmann the thermodynamic average, denoted here $\langle d_{O-O} \rangle$, is simply the usual average over all values d_{O-O} computed from the atomic positions in the simulated system. If a total of M such values d^i_{O-O} can be computed from the configurations, Equation (1) applies.

$$\langle d_{O-O} \rangle = \frac{1}{M} \sum_i^M d^i_{O-O} \tag{1}$$

[Top of Page]

2. Fundamental Approximations

At the energies relevant for most of biology, chemistry, and engineering, matter can be considered as consisting of molecules which can undergo changes (chemical reactions). The molecules themselves consist of atoms, which in turn consist of nuclei and electrons. Such particles of molecular, atomic, or subatomic dimensions are, in principle, subject to quantum mechanics, as expressed, e.g., through the Schrödinger equation for the wavefunction, rather than to classical mechanics, i.e., the Newton equation of motion. In most cases of interest here, however, approximations can be used to simplify the problem. In the Born – Oppenheimer approximation [3], it is recognized that the mass of the electrons is much smaller than the mass of the nuclei. The problem concerning the electrons and that concerning the nuclei can thus be treated separately. The energies holding together the atoms to form molecules as well as the energies holding together molecules, e.g., in liquids or solids, can thus be computed by solving quantum mechanical Schrödinger equations only for the electrons. It is one of the main purposes of quantum chemistry to devise methods, usually numerical, to tackle this complex problem for molecular or supramolecular systems [3-5]. Unfortunately, even with the largest computers available in the foreseeable future, it will not be possible to elaborate these energies for extended systems consisting of large disordered assemblies of molecules. Yet, according to what has been said above, in order to do statistical mechanics, as many molecular arrangements as possible compatible with a given set of constraints must be constructed, and to do this, the energies are needed.

The interaction energies between particles (individual atoms, atoms in molecules, groups of atoms, molecules themselves), which can be obtained, in principle, from the solution of the Schrödinger equations for the electrons, can then be expressed as potential functions, or potentials. They are usually written as functions of the positions (and, if needed, orientations) of the particles. Simplifying assumptions as to the mathematical form of these potentials are introduced at this level (see Chap. Interaction Models).

The second major approximation is that classical dynamics is used for the nuclei themselves, i.e., the atoms and molecules, rather than quantum mechanics, to construct the ensemble of configurations. The heavier the masses involved, the weaker the interactions, and the higher the temperature, the better this approximation will be. Equivalently, the smaller the so-called thermal De Broglie wavelength $\Lambda = \sqrt{\frac{h^2}{2\pi m k_B T}}$ associated with a particle of mass m at temperature T compared to a characteristic length (in a gas, e.g., on the order of the mean free path), the better the classical approximation [6]. At room temperature, the classical dynamics become a limitation only for the lightest atoms (e.g. hydrogen, helium) when they are strongly bonded.

Before we proceed we should note here that in selected cases, essentially for small systems for which not too large a number of configurations is sufficient, the quantum mechanical problem for the electrons can be solved (usually in a simplified approach) simultaneously with the generation of the configurations. Thus, no explicit interaction model must be specified, although the quantum mechanical calculations usually involve a number of adjustable quantities which are empirically chosen. One popular approach is Car – Parrinello MD (CPMD) [7]. Here, electrons belonging to all particles present in the system are treated, in the density-functional theory (DFT) approach, through equations of motion. This method is not free of arbitrariness, e.g., in the choice of the so-called functional; it is extremely CPU-intensive and is thus restricted to very small systems (a few tens of small molecules) which can be explored in a short time (a few picoseconds). In other approaches, the quantum mechanical equations are retained only for a small part of the system, while the larger part is described classically. The system is usually divided up geometrically into a smaller region which is treated quantum mechanically (by the ab initio, DFT, or, mostly, semi-empirical approach) and a larger one in which the interactions are described by an interaction model (force field). Coupling these two regions together is problematic and again involves choices and approximations. These so-called quantum-classical simulations are mostly outside the scope of this article, but we stress that if systems with real chemical reactions are to be studied approaches like the ones sketched here must be used.

Besides the two fundamental approximations (classical mechanics and interactions via potentials), several more are needed in order to study most systems of interest, mainly in two areas: The way the interactions are described, and the approximations needed to mimic extended systems with finite computational resources. The following sections are devoted to these aspects. In the generally available simulation packages (freeware, shareware, commercial) these approximations are often dealt with in a way somewhat hidden from the user. It is not the purpose of this article to evaluate the various approaches proposed in these packages and to give recommendations. We believe, however, that users of such software should be aware of the approximations involved and of their consequences: All results should always be

critically evaluated in the light of the approximations made.

[Top of Page]

3. Interaction Models

The interaction model, often called force field, is, beyond all further “technical” approximations to be discussed below, the crucial input to a classical MD simulation. Within the restrictions in time and space inherent to the technique, the simulation is supposed to yield the “exact” results associated with a given interaction model. In other words, failure of the simulation to reproduce a given observable in a given system should be ultimately ascribable to deficiencies in the interaction model.

Two different philosophies can be distinguished: In the first, MD simulations are used to explore systematically the statistical mechanics of certain types of interaction models, usually highly idealized simple ones (e.g., hard spheres or Lennard–Jones spheres [8], see Eq. 7). Such simulations thus yield in some sense the “exact” result for the idealized model. The second approach attempts to reproduce as precisely as possible the properties of specific chemical systems. Water has been, for obvious reasons, a frequent target of such studies [9]. In the latter case, it is a reasonable and efficient approximation to write the total potential energy of a molecular system U^{total} as a sum of terms (Eq. 2)

$$U^{\text{total}} = \sum_{\text{all moleculle}} U^{(1)} + \sum_{\text{all pairs of molecules}} U^{(2)} + \sum_{\text{all triples of molecules}} U^{(3)} + \cdots \tag{2}$$

where $U^{(1)}$ is an energy associated with a single molecule, e.g., its intramolecular energy (see below) or its energy in an external field. For small molecules, this term is often nearly constant and thus omitted in the simulations. $U^{(2)}$ is the interaction energy between pairs of molecules. In most cases, only terms of this type are used in the simulations. $U^{(3)}$ is an additional interaction energy that depends simultaneously on (e.g., the positions of) three molecules. Figure 1 illustrates the second and third terms of this series of approximations.

Figure 1. Left: Interaction $U^{(2)}$ (solid lines) between two molecules represented as circles. Middle: Interactions $U^{(2)}$ between three molecules; the total energy is the sum of the three $U^{(2)}$ terms, each of which is the same as in the two-body case, i.e., each interaction is deemed to be independent of the presence or absence of the third molecule. Right: Interactions between three molecules as the sum of the two-body terms $U^{(2)}$ plus a modifying three-body term $U^{(3)}$ (dashed circle).

In a system of N molecules, there are N terms of the first kind; the number of terms of the second kind increases as $N(N–1)/2 \propto N^2$, and the number of terms of the third kind increases proportionally to $N(N–1)(N–2)/6 \propto N^3$. Even though not all terms need to be evaluated, these proportionalities govern the amount of computer time needed to evaluate the interactions in a simulation.

3.1. Intramolecular Interactions

Intramolecular interactions are usually stronger than intermolecular interactions. If this is the case, the molecular geometry can be assumed not to change much when a molecule interacts with its neighbors. Since the intramolecular energy will then be constant, it can be neglected, even though it may be large: A rigid body is then a suitable molecular model. This works in some instances for small molecules. In many cases, however, and in particular for large molecules, the energetics involved with changes in the molecular geometry must be included in the simulation. Since it is chemically intuitive, the intramolecular energy is often expressed as a power series in internal displacement coordinates ρ [10] (Eq. 3).

$$\begin{aligned} U^{(1)} &= \sum_{\text{all pairs}(l,m)\text{of coordinates}} k_{lm} \cdot \rho_l \cdot \rho_m + \sum_{\text{all triples}(l,m,n)\text{of coordinates}} k_{lmn} \cdot \rho_l \cdot \rho_m \cdot \rho_n \\ &+ \sum_{\text{all quadruples}(l,m,n,o)\text{of coordinates}} k_{lmno} \cdot \rho_l \cdot \rho_m \cdot \rho_n \cdot \rho_o + \cdots \end{aligned} \tag{3}$$

Examples of such internal displacement coordinates are the bond stretch $\rho = r_{ij} – r_{ij}^0$, where r_{ij} denotes the module of the distance between two atoms in a molecule and r_{ij}^0 its equilibrium value, and the angle bend $\rho = \alpha_{ijk} – \alpha_{ijk}^0$, where α_{ijk} denotes the angle spanned by particles i and k at the position of particle j. The constants k_{lm} are called the force constants; their definition in the literature varies due to the inclusion or omission of the factor of 1/2 originating from the definition of the power series. The same holds for the higher constants k_{lmn} and k_{lmno}. Note that this definition of $U^{(1)}$ rests upon the notion of equilibrium geometry, i.e., it is suitable for molecules for which only one conformer must be considered.

In the first term of Equation (3), the terms with $\rho_l = \rho_m$ are called diagonal terms. Nondiagonal terms ($l \neq m$) are usually included only if the corresponding coordinates ρ_l, ρ_m share a common particle (i or j). The higher terms in Equation (3), usually called (mechanical) anharmonicities, are also not frequently used. In some cases, however, they cannot be omitted since they play an important role in the redistribution of energy in the system. Another frequently used approximation for anharmonic intramolecular bond stretches is the Morse potential (Eq. 4) [11]

$$U^{\text{Morse}}(r_{ij}) = D \cdot \left(1 - e^{-a\left(r_{ij} - r_{ij}^{o}\right)}\right)^2 \tag{4}$$

where D is the depth of the potential well, and a a constant.

We note again that the number of operations needed in a simulation to evaluate Equation (3) is proportional to *N*, the number of molecules. Thus, the computational weight of these terms is usually not very large, in contrast to the intermolecular terms (see Section Intermolecular Interactions). More details about the modeling of intramolecular interactions, e.g., further definitions of internal coordinates, information on the quantum mechanical background, and potential functions suitable for molecules with several conformers can be found in ⟶ Molecular Modeling.

3.2. Intermolecular Interactions

3.2.1. *Pair Potentials*

In most cases the intermolecular interactions are expressed in terms of potentials of type $U^{(2)}$ only (see Eq. 2). A frequently used approach consists of writing this interaction potential as a sum of so-called site–site pair potentials (Eq. 5)

$$U^{(2)} = \sum_{\substack{\text{all sites } i \\ \text{on molecule } \alpha}} \sum_{\substack{\text{all sites } j \\ \text{on molecule } \beta}} U_{ij}(r_{ij}) \tag{5}$$

In many instances, “site” is identified with “atom position”, but if molecules are described as rigid entities, other points on the molecular frame can be used as sites. Examples of such site–site pair potentials are the electrostatic (Coulomb) interaction between two partial electric charges *q* (Eq. 6)

$$U^{\text{Coulomb}}(r_{ij}) = C \cdot \frac{q_i \cdot q_j}{r_{ij}} \tag{6}$$

where *C* is a constant depending on the energy unit used. We note already here that special techniques such as Ewald summations or reaction field techniques [12], which are computationally expensive, must be used to evaluate Coulomb terms in the periodic arrangements that are mostly used to represent extended systems. This is due to the long range ($U \propto r^{-1}$) of this interaction.

Shorter range repulsive and attractive (van der Waals-type) interactions are often approximated by the Lennard–Jones (LJ or 12-6) potential, which contains two constants ε and σ (or *A* and *B*; Eq. 7).

$$U^{\text{Lennard-Jones}}(r_{ij}) = 4 \cdot \varepsilon_{ij} \cdot \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \tag{7}$$

The constants ε and σ used in the first notation represent the depth of the potential well and the value of the argument *r* for which $U^{\text{Lennard–Jones}} = 0$, respectively. As an illustration of the typical order of magnitude for the two interaction types mentioned above, Figure 2 shows a Lennard–Jones potential and a Coulomb potential between two small partial charges of opposite sign, typical for those on nonpolar molecular groups. The examples given below will show that about ten times higher partial charges are used to represent polar groups, and even higher ones for ions.

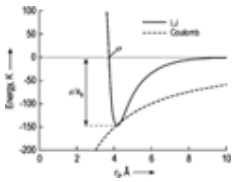

Figure 2. Typical interaction potentials as a function of distance. Lennard–Jones (LJ) potential between two methane (CH_4) molecules carrying one site each (ε/k_B = 148.6 K, σ = 3.728 Å, 1 Å = 10^{-10} m) (effective potential in the united-atom approximation). For comparison: Electrostatic Coulomb potential between two opposite charges of $q = 0.06\ |e|$, where |*e*| is the elementary charge. The energy is given in Kelvin, i.e., divided by k_B, since the thermal energy is in some way the measure of all things in a thermodynamic system.

Since the number of parameters needed to specify the interactions (Eqs. 6 and 7) increases rapidly when several types of molecules must be described (see Example 2. in Section Intermolecular Interactions), empirical rules, called combination rules, are often used to construct the LJ parameters ε and σ for interactions between unlike sites (*ij*) from those for interactions between like sites (*ii* and *jj*). There are several sets of such rules, the most frequently used of which are due to Kong [13] and Lorentz–Berthelot [14]. The Lorentz – Berthelot rule is given here as an example (Eq. 8).

$$\sigma_{ij} = \frac{1}{2} \cdot (\sigma_{ii} + \sigma_{jj}), \quad \varepsilon_{ij} = \sqrt{\varepsilon_{ii} \cdot \varepsilon_{jj}} \tag{8}$$

Another way to simplify the expressions for the intermolecular energies is to attempt to reduce the number of sites used to represent a group of atoms or a molecule. The CH_n-groups in organic molecules are thus often represented by just one site instead of several (e.g., one site on each atom). This is known as the united-atom approximation. It may work if the group is not too polar. Figure 2 shows the interaction potential resulting between two CH_4 molecules, each of which is represented by just one site. We note that this potential is spherically symmetric, and it thus neglects the exact “shape” of the molecule. Typical simulation results for this system are discussed in Examples.

Many other functional forms have been proposed besides the Lennard – Jones ansatz to describe intermolecular interactions of type $U^{(2)}$. The powers 12 and 6 in the LJ-ansatz (Eq. 7) have thus been modified and/or written as functions of *r*. In other instances (e.g., in the Buckingham potential) the inverse power of *r* describing the repulsive part of the potential has been replaced by an exponential function. Alternative forms of molecule–molecule potentials may involve functions of relative molecular orientations instead of sums of site–site potentials. The Gay–Berne model [15], frequently used in simulations of liquid-crystalline systems, again uses a LJ-type form, but with parameters dependent on the relative orientation of two molecules. A detailed discussion of intermolecular interactions in simple systems can be found in, e.g. [16].

3.2.2. *Three-Body and Higher Order Intermolecular Potentials*

Since they are difficult to establish and computationally expensive, three-body ($U^{(3)}$) and higher terms are not often used in simulations. This is an area of active research, and we mention here only two examples among the many approaches.

The three-body Axilrod – Teller interactions [17] (Eq. 9)

$$U^{\text{Axilrod-Teller}} = \frac{\gamma \cdot \left(1 + 3 \cdot \cos\theta_{ij} \cdot \cos\theta_{jk} \cdot \cos\theta_{ki}\right)}{r_{ij} \cdot r_{jk} \cdot r_{ki}} \tag{9}$$

where *i*, *j*, and *k* are three particles, similar to Figure 1, *r* their distances, and θ the angles. This model potential contains only one constant, γ, which can be related to the molecular polarizability [3].

The Sutton – Chen [18] model, which has been used successfully, e.g., in studies of metals and surfaces, is formally similar to an LJ-ansatz with generalized exponents *n* and *m*. The second term, however, is replaced by an average over all interaction partners *j*. This is thus a true many-body potential, and the total potential energy of the *N*-particle system is written as Equation (10)

$$U^{\text{Sutton-Chen}} = \varepsilon \cdot \sum_{i}^{N} \left(\frac{1}{2} \sum_{i \neq j} \left(\frac{a}{r_{ij}} \right)^{n} - b \cdot \sqrt{\sum_{i \neq j} \left(\frac{a}{r_{ij}} \right)^{m}} \right) \tag{10}$$

Five constants, ε, *a*, *b*, *n*, and *m* must be specified here.

An important physical effect neglected in the pair potential approach discussed up to here is the mutual electronic polarization of molecules. There have been many attempts to devise workable schemes allowing to treat the polarization $\vec{P}$ of a molecule efficiently and as exactly as possible in a simulation. $\vec{P}$ can be seen as the additional dipole moment created on a molecule by the electric field $\vec{E}$ generated by its neighbors. In the picture of Figure 1 this would mean that the intermolecular interaction between two molecules, denoted there $U^{(2)}$, is modified by the presence of a third, fourth, fifth molecule in the neighborhood. In the notation of electrodynamics [19], one has $\vec{P} = \overleftrightarrow{\alpha} \cdot \vec{E}$, where $\overleftrightarrow{\alpha}$ is the molecular polarizability tensor. Many difficulties arise: $\overleftrightarrow{\alpha}$ is often not well known; if it is, it is not clear with which site of a molecule it should be identified or how it should be distributed among the possible locations on the molecular frame. Furthermore, if there is more than one polarizable species in the system, an iterative self-consistent procedure is in principle needed to evaluate the energies resulting from this effect: The field generated by the neighborhood polarizes molecule *i*, say, but the changing electric moments (usually only the dipole moment is considered) of molecule *i* will change the electric field experienced by the neighbors, hence their polarization, hence the field acting on *i*, etc., etc. Several sophisticated approximate methods have been proposed to treat this problem (see, e.g., [20], [21]); they cannot be described here. We note, however, that if intramolecular potentials are specified (flexible molecules) some effective mutual polarization will arise from the mutual geometrical distortions of the molecules, e.g., if they lead to distance changes between point charges. In another particularly simple approach, called the fluctuating charge model, the partial electric charges q_i are allowed to vary, e.g., as explicit functions of some distances or angles in the system.

Finally, we note that in certain instances the interaction energies and/or their derivatives, the forces, need not be expressed explicitly as mathematical functions, but can simply be tabulated in one- or multidimensional arrays. Interpolation schemes or splines are then used in the simulations to compute the relevant quantities.

3.3. Example: Intermolecular Potentials for an Aqueous Salt Solution

We stress again that the choice of the interaction model is crucial. It involves the choice of a proper mathematical representation and the choice of all the required parameters. In particular, the choice of the number of sites used to express the intermolecular interactions is crucial. All these choices are largely based on experience and result from a compromise between a representation of the interactions that is as accurate as possible and the resulting computational expenditure required to generate a sufficient number of configurations (see Section Method). An example will illustrate this:

Example 2. A simulation of a solution of a 1:1 salt, say NaCl, in water. The site–site pair potential approach is used for all intermolecular interactions. Rigid molecules (i.e. no intramolecular potentials need to be specified) are assumed and there is no external field.

We consider a system consisting of three species, the solvent (H_2O), cations (Na^+), and anions (Cl^-). The simplest reasonable representation of the water molecules needs three sites (say the atoms O and H), while one site suffices for these simple ions. There are thus four types of sites, leading to (4 × 5)/2 = 10 site–site intermolecular pair potentials $U^{(2)}$ (see Equation 2), which can be arranged in an upper triangular matrix for better visibility:

$U^{(2)\text{O–O}}$	$U^{(2)\text{O–H}}$	$U^{(2)\text{O–cation}}$	$U^{(2)\text{O–anion}}$
	$U^{(2)\text{H–H}}$	$U^{(2)\text{H–cation}}$	$U^{(2)\text{H–anion}}$
		$U^{(2)\text{cation-cation}}$	$U^{(2)\text{cation-anion}}$
			$U^{(2)\text{anion-anion}}$

To write each of these 10 potentials as a sum of a Coulomb interaction (Eq. 6) between electric charges and a Lennard – Jones term (Eq. 7), 20 Lennard – Jones constants (ε and σ, or *A* and *B*) plus four electric partial charges are needed. In this case, it can be assumed that the entire ionic charge of ±|*e*| resides on the ions. The water molecules being neutral, one has $q_O = -2q_H$; there is thus only one parameter to determine, say q_O, which is of the order of magnitude of $q_O \approx -2/3|e|$. As far as the Lennard–Jones parameters are concerned, assuming combination rules (e.g., those in Eq. 8) would reduce the number of parameters to 2 × 4 = 8 for the potentials $U^{(2)}$ in the diagonal of the matrix above, the parameters of

the off-diagonal $U^{(2)}$ resulting from those of the diagonal terms. We note in passing that the notion of "hydrogen bond" does not appear explicitly in the model. The typical features associated with these bonds in terms of structural and dynamical properties of the liquid arise from the interplay between the pair potentials. In any case, it is easily seen that the number of parameters which must be known increases very quickly with increasing complexity of the system, i.e., with the number of species in the system and with the sophistication (or complexity) of the description.

3.4. Data for Interaction Models

Details of the various approaches used to determine consistent sets of parameters for interaction models are beyond the scope of this article. There are essentially three sources of information used for this purpose:

1. Quantum chemistry: This is the "in principle" approach outlined in Chapter Fundamental Approximations. For practical aspects (⟶ Molecular Modeling – Quantum Mechanical Modeling). It will suffice here to say that quantum mechanics, at various levels of approximation (e.g. ab initio, semi-empirical, etc.), can be used to determine, as a function of the relative arrangements of molecules, intermolecular interaction energies in ways very similar to the determination of intramolecular ones. One then endeavors to fit these energies with expressions of type $U^{(2)}$, $U^{(3)}$, etc. We note, in particular, that although the notion of partial charge is not inherent to quantum chemistry, most quantum chemical program packages provide algorithms which help in the determination of these quantities by suitable integrations over the nuclear charges and the wave functions or by fitting procedures.
2. Experiments: Thermodynamic data have frequently been used to gauge the comparative depths of potential wells. Very careful work has, e.g., been carried out to reproduce IR and Raman spectra of many molecules by "harmonic force fields" (essentially the first term in Eq. 3) or higher order expansions. These provide valuable input for intramolecular model potentials. Results from scattering experiments can provide insight into molecular shapes, i.e., essentially the repulsive part of the interaction potential.
3. Simulations themselves: Simulations are often used in a sort of trial and error procedure to refine model potentials, i.e., to adjust potential parameters. Again, this is often done with thermodynamic data. Model potentials have thus been adjusted to yield the proper pressure–density dependence. Self-diffusion coefficients have also been used for such procedures. Some systematic procedures have been proposed recently [22]. Note that there is in general no guarantee that the solution thus found for the interaction model is unique.

In all cases, "true" pair potentials, as constructed, e.g., from quantum mechanical calculations of pairs of molecules, must be distinguished from "effective" potentials, in which averages over the higher order terms are in some way worked into the $U^{(2)}$ term of Eq. 2. Many models, e.g., have a distribution of electric charges (Eq. 6) leading to enhanced dipole moments compared to the gas-phase value in order to account for the average polarization of molecules by their neighbors in a condensed phase. Consequently, such models are in principle valid only for a restricted range of densities and temperatures.

[Top of Page]

4. Approximations for the Simulations

Molecular dynamics simulations are mostly used to study condensed phases, homogeneous or inhomogeneous. They are also very well adapted to studies of interfaces. For the sake of simplicity we shall, however, focus here on simulations of simple homogeneous systems, say a liquid under ambient conditions. The question arises: How many particles does one need for the typical properties of a liquid to be observable? Or, in other words, how many particles/molecules need to be included in the simulation so that the properties determined from this simulation reflect those of a bulk liquid, and not of some sort of cluster, aggregate, or droplet?

Assume for the present qualitative argument the particles to be arranged in a cube with an edge length equal to a multiple of the particle dimension (not worrying what exactly this dimension might be). We consider cubes of 3 × 3 × 3 = 27, 4 × 4 × 4 = 64, 125, 216, 343, . . ., 15 × 15 × 15 = 3375, . . ., particles (the number 3375 being already relatively high for a simulation with anything but the simplest interaction model, although simulations with many more particles have been reported in special cases). Assuming furthermore that the influence of the surface extends about three molecular layers deep into the system, at best (15–6)(15–6)(15–6) = 729 particles would display bulklike properties (i.e., unperturbed by the surface). Thus, in this example, only 729/3375 or 21 % of the particles included in the simulation would contribute to the result. Placing the particles in a sphere would of course improve this ratio, but not decisively.

To avoid this unaffordable inefficiency, the system to be investigated is replaced by a periodic one through the introduction of periodic boundary conditions. The simplest case of cubic periodic boundary conditions is illustrated in Figure 3. The construction is equivalent to replacing the *N*-particle system by a system of *N* cubic primitive lattices: Each particle *i* forms with its image particles *i′* in the neighboring image boxes a cubic primitive lattice. Together with the minimum-distance convention, which stipulates that the interaction is computed between the two closest partners, be they two particles in the basic box or a particle in the basic box and one in a neighboring image box, two advantages are gained:

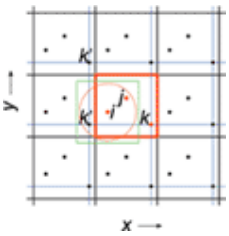

Figure 3. A system with cubic periodic boundary conditions containing three particles *i*, *j*, and *k* (in red). The basic cube, usually called box, i.e., the volume containing the *N* (here 3) particles of the simulation, is highlighted in red. It is surrounded by 26 (in 2 dimensions 8) "image boxes" (in black), which themselves are surrounded by further boxes. Each image box contains periodic images (in black) of all particles in the basic box. The green box is centered on particle *i*. It can be seen that it contains particle *j* but instead of particle *k* it contains an image *k′*. Under the minimum-distance convention particles *j* and *k′* are the interaction partners of particle *i*. The orange sphere (circle in 2 dimensions) centered on particle *i* indicates that under certain conditions the interactions of particle *i* with particles inside the green box, but outside of this sphere, can be neglected. The radius of the sphere is known as the cutoff radius r_{cut}. If applicable, this procedure can save a lot of computer time, the neglected volume (cube vs sphere) being roughly 50 %.

1. There are no “surfaces” in the system, since each particle is, owing to this construction, at the center of all its interaction partners (green box in Figure 3).
2. The density in the basic box (and thus in the system) remains constant. If a particle leaves the basic box on one side during a simulation, it becomes an image particle. An image particle enters simultaneously on the other side of the box and becomes the particle.

These procedures are not without incidence on some properties computed from an ensemble simulated in this way. In some instances, the correction is easy (maybe even trivial), in others, it is difficult even to estimate its influence. As a rule of thumb, the size of the simulation box and, related through the density of the system, the number of particles, should be chosen in such a way that spatial correlations (expressed, e.g., through the radial pair distribution function; see Chap. Applications) decay over less than half the box length (i.e., from a central particle to the edge). Furthermore, in systems where collective motions (e.g., phonons) play an important role, one should not forget that the box will limit the wavelengths associated with the phenomenon. In mixtures, e.g., solutions, the desired concentration may also set limits as to the minimum number of particles that must be included.

As indicated in Figure 3, the calculations can, under certain circumstances (i.e., when the range of the interactions is short) be accelerated even more by truncating the interactions beyond a cutoff radius r_{cut}, i.e., the potential used in the simulation is given by Equation (11).

$$U^{\text{simulation}} = \begin{cases} U^{\text{model}} & \text{if } r < r_{\text{cut}} \\ 0 & \text{else} \end{cases} \tag{11}$$

There are several ways to implement this truncation scheme [12]. Note that truncation schemes can generate discontinuities in the forces which must be dealt with.

[Top of Page]

5. Constructing the Ensemble

5.1. Method

The task is now to construct a sufficient number of microscopic configurations, or microstates, compatible with the interactions and the macroscopic constraints. We consider only systems set up as described above, i.e., with a fixed number of particles *N* in a fixed (through the periodic boundary conditions) volume *V*. We assume that it is possible to construct one such microstate, i.e., to find positions and velocities for all particles such that the total energy of the system *E* (Eq. 12)

$$E = U^{\text{total}} + E^{\text{kin}} \tag{12}$$

where E^{kin} is the kinetic energy, has a desired value. A series of new microstates, constrained by *N* = const., *V* = const., and *E* = const., can be generated by solving the Newton equation (or any equivalent set of equations) for all particles under the influence of their mutual forces. We shall call this an NVE simulation. Written here, for simplicity, for *N* point masses m_i, i.e., without rotations, these are Equations (13).

$$\begin{aligned} m_1 \ddot{\vec{r}}_1 &= \vec{F}_1\left(\vec{r}_1, \vec{r}_2, \vec{r}_3, \ldots, \vec{r}_N\right) \\ m_2 \ddot{\vec{r}}_2 &= \vec{F}_2\left(\vec{r}_1, \vec{r}_2, \vec{r}_3, \ldots, \vec{r}_N\right) \\ &\vdots \\ m_N \ddot{\vec{r}}_N &= \vec{F}_N\left(\vec{r}_1, \vec{r}_2, \vec{r}_3, \ldots, \vec{r}_N\right) \end{aligned} \tag{13}$$

They are coupled through their right-hand sides, i.e., through the forces. The force acting on mass *i* depends, in general, on the position of mass *i* itself and on the positions of all other masses *j* through Equation 14.

$$\vec{F}_i = -\nabla_i U^{\text{total}} \tag{14}$$

Note that U^{total} (Eq. 2) is given as a function of the site positions, which may or may not coincide with the positions of the masses. In molecules or groups of atoms treated as rigid bodies, the equations become more complex [12], but the principle remains the same.

Additionally to the constraints (NVE), the microstates constructed in this way fulfill the additional constraint of constant total momentum (Eq. 15).

$$\vec{p} = \sum\left(m_i \vec{v}_i\right) = \text{const.} \tag{15}$$

This constant is usually set equal to zero, which facilitates the subsequent evaluation procedures (see Section Simple Averages). The configurations thus generated constitute a particular sample of the microcanonical ensemble of statistical mechanics (see, e.g., [6] or any other book on statistical mechanics) and averages can be computed over this sample, as exemplified in Equation (1).

Except for very few specific cases it is not possible to find a closed, analytical solution of Equations (13). They are thus solved numerically, and several numerical schemes have been devised for this purpose. Without going into details, all these integration schemes can be thought of as procedures in which the following steps are iterated:

1. Knowing the positions and velocities of all particles (sites), calculate all forces (and torques if rigid groups are considered) acting in the system
2. For a short time step δt, allow the particles (a) to move according to their velocities and (b) accelerate according to the forces acting on them
3. Determine new positions and new velocities;
4. Go to step 1

The microstates, or configurations (which we shall denote as $\mathscr{R}$; $\mathscr{R}$ = all positions and velocities of all particles) are thus points on the trajectory of the system evolving with time according to the Newton equation. The generated ensemble can thus be written as a series of configurations ordered according to time and starting at a time origin t_0:

$$\{\mathcal{R}(t_0), \mathcal{R}(t_0+\delta t), \mathcal{R}(t_0+2\delta t), \ldots\ldots\ldots, \mathcal{R}(t_0+M\delta t)\}$$

if *M* integration steps have been carried out. In simulation parlance the term “history” of the system is often used for the trajectory. The availability of the time evolution of the system is the main difference between MD and MC simulations.

The length of the time step δt must be chosen carefully. It should be as large as possible in order to sample the system efficiently (minimize computational costs) but short enough to ensure correct integration of the equations of motion. In NVE simulations, the choice of δt can be checked by making sure that the fluctuations in total energy (which should be zero since *E* = const., Eq. 12) are small compared to real fluctuations in the system, e.g., between kinetic and potential energy. In terms of dynamics, the time-step size must be at least one order of magnitude smaller (say a factor of 20–50) than the period of the fastest motion occurring in the system (e.g., molecular O–H or C–H vibrations, Debye frequency in a solid, etc.). If there are very different motional regimes in the system, it is sometimes possible to use a multiple-time-step method in which the interactions of the slowly fluctuating parts are evaluated less often than those of the fast ones.

We have dealt here explicitly only with the generation of the NVE ensemble. Most simulation packages offer the possibility to generate other ensembles, mostly NVT or NpT, i.e., with constant temperature *T* and/or constant pressure *p* instead of constant energy *E* and/or constant volume *V*. There are various ways to proceed, e.g., by solving additional differential equations associated with “thermostats” and “barostats” or by adding “virtual particles” that act as heat reservoirs; see [23] for details. Both methods may occasionally lead to artefacts [24]. We note in particular that in both cases the dynamics are no longer strictly Newtonian, which may have consequences for the dynamical properties computed from such simulations. However, using these methods for equilibration can facilitate this step, and so an equilibration procedure (see below) using these methods is often combined with a final production run in the NVE ensemble.

5.2. Setting up a Simulation

It was assumed above that a suitable configuration is available to start the simulation. It is by no means obvious how to generate such a configuration but a procedure similar to that outlined in Figure 4 often leads to success.

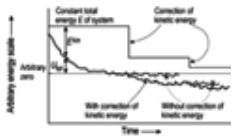

Figure 4. Schematic equilibration procedure to generate a starting configuration for an equilibrium simulation under NVE conditions.

The particles are placed in the simulation box as “reasonably” as possible, e.g., randomly but with reasonable distances and orientations, or on a lattice or distorted lattice. Velocities are randomly assigned or simply set equal to zero. The integration is started. Assuming the numerical integration works perfectly, the total energy of the system will remain constant. Since the selected configuration will most probably have a surplus of potential energy (a slightly too short interparticle distance on a repulsive part of a Lennard–Jones potential (Eq. 7) suffices) conversion of potential energy to kinetic energy will occur, and the temperature of the system will rise. If the interaction potentials are sufficiently anharmonic, the surplus energy will be redistributed among all degrees of freedom, and the system will converge toward a regime where there are only fluctuations between the two energies, but no systematic drift. (In contrast, a system exhibiting motions of the normal-mode type will not show this behavior.) However, the average kinetic energy (i.e., the system temperature) is now usually much higher than desired. Kinetic energy must thus be removed, either “instantaneously” (as in Figure 4) by scaling the velocities, or more gradually through thermostatting procedures. In any case, since average kinetic and potential energies of a system are interrelated, there will be a relaxation following this adjustment. However, through a series of such adjustments the system can be brought to a state with a total energy compatible with the desired average temperature. All configurations generated during this equilibration procedure are then discarded, except the final one, which is the true initial configuration of the simulation.

It is generally not easy to determine whether one has really reached equilibrium. Necessary conditions are that there should be no drift with time of the kinetic or potential energies, that the particle velocities should follow approximately the Maxwell–Boltzmann distribution, and that the time derivatives of various quantities computed from the ensemble should be zero. However, all this does not guarantee that the system is not trapped in a metastable state.

Equilibration is usually not a problem for homogeneous liquids consisting of small molecules under conditions of temperature and density where the molecular mobilities are not too low (say compared to the time accessible in a simulation, see Figure 5). In mixtures, the equilibration may be more difficult. It becomes a major problem in simulations of systems of large molecules such as polymers or biomolecules.

Another test is to use different starting configurations and equilibration procedures. Once equilibrium has been reached, properties computed over the equilibrium ensemble should of course not depend on the way the equilibrium was reached.

Many simulation programs report observables (such as the system temperature, kinetic energy, potential energy, momenta, etc., and also partial and total averages of these quantities) “on the fly”, i.e., while the simulation is still in progress. While this information is very useful for monitoring the advancement of a simulation and detecting problems, one will eventually want to extract much more information from the simulated ensemble. It is thus necessary to store the configurations on some storage device at time intervals short enough for a subsequent statistical evaluation. This can result in large files. As an example, for a system of 1000 argon atoms, storing just 10 000 configurations (positions and velocities) in single precision (unformatted) would take 1000 (particles) × 10 000 (configurations) × 6 ($x/y/z/v_x/v_y/v_z$) × 4 bytes = 240 Mbyte of storage space.

[Top of Page]

6. Obtaining Results

Once the sample of the thermodynamic ensemble has been constructed, statistical methods can be used to extract the macroscopic observables of the system. This is where a very critical assessment is needed whether the ensemble of simulated configurations is a sufficiently representative sample of the thermodynamic ensemble. Questions to be answered include: Has sufficient decorrelation from the initial configuration been achieved? Could the system be trapped in a metastable state? There are no rigorous methods to tackle these questions, and trial and error methods as well as experience are most often called upon here. Analytical methods [25] can be used in some cases to gauge the reliability of the computed averages in a strictly statistical sense.

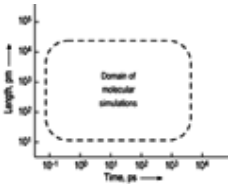

Figure 5. Typical space and time domain accessible to molecular dynamics; see text for typical physical phenomena occurring in this domain.

In summary, taking into account the approximations and limitations described above, Figure 5 shows roughly the space–time window accessible to MD computer simulations with today's (ca. 2004) programs and computers. The lower left corner of the graph illustrates the time–space domain of typical fast (in a chemical sense) phenomena such as the electronic transitions observed in optical spectroscopy. The time domain associated with molecular vibrations ranges from femto- to picoseconds; molecular reorientations in liquids have correlation times of several picoseconds. These phenomena can be probed, e.g., by IR and Raman spectroscopy in classical and time-resolved modes, as well as by NMR spectroscopy and dielectric techniques. Longer times are associated with motions of large molecules and with collective motions of groups of smaller ones. The length scale extends from the Ångström range, typical of interatomic distances, accessible experimentally by scattering methods (X-rays, neutrons, etc.), to the nanometer range, which characterizes medium-sized assemblies of molecules. The hydrodynamic regime starts at the top right corner of the graph, it is accessible through other types of simulations (see Chap. Applications).

Two kinds of averages are usually distinguished: Simple averages over all available configurations $\mathscr{R}$, and averages in which the order of the configurations, i.e., the time evolution of the system, is taken into account. In the equilibrium ensemble, this is most often done by computing so-called time-correlation functions. We note here, without going into detail, that these time-correlation functions and the spatial correlations briefly discussed below can be seen as limiting cases of a general space–time function [6].

6.1. Simple Averages

The average value of a quantity A, denoted $\langle A \rangle$, is computed as Equation (16)

$$\langle A \rangle = \frac{1}{M} \cdot \sum_{i}^{M} A(\mathcal{R}_i) \tag{16}$$

where A is a quantity that can be determined from the particle positions, velocities, (and possibly forces) contained in the system configurations R_i. Equation (1) is an example of this procedure. Another example follows from the equipartition theorem of statistical mechanics, written in Equation (17) for a system of N masses

$$\frac{3N-3}{2} \cdot k_B \cdot \langle T \rangle = \langle E^{kin} \rangle = \frac{1}{2} \cdot \sum_{i}^{N} m_i v_i^2 \tag{17}$$

where k_B is the Boltzmann constant, $\langle E^{kin} \rangle$ is the average kinetic energy and $\langle T \rangle$ the average temperature. The term $3N$–3 is the number of degrees of freedom in this system. The three degrees of freedom corresponding to the movement of the center of mass must be removed by virtue of Equation (15). In general, for systems with constraints or with molecules represented by rigid or semirigid entities, the real total number of degrees of freedom must be entered into Equation (17) instead of $3N$–3. Examples of such averages are, besides $\langle T \rangle$ in Equation (17), the average potential energy $\langle U \rangle$, the average pressure $\langle p \rangle$, and many other thermodynamic quantities which can be directly related to experimental values. Thermodynamic quantities involving derivatives can be obtained from fluctuations (e.g., heat capacities, for which quantum corrections may have to be included) or by more advanced statistical mechanical treatments [26-28]. Spatial correlations like the radial pair distribution function (RDF) $g(r)$ or its analogue, the radial pair correlation function $h(r) = g(r)$–1, are routinely computed. These functions are suitably normalized histograms over interparticle distances between pairs of molecules; they are related to the so-called potential of mean forces (PMF) and are often used in thermodynamics calculations [6]. They are furthermore related through Fourier transformations to the scattering functions obtained in X-ray

or neutron scattering experiments [29].

We note that in Equation (16) the result will not depend on the order of the configurations $\mathcal{R}_i$. The fact that the $\mathcal{R}_i$ were constructed using the Newton equation is thus of no consequence. Quantities like $\langle A \rangle$ are thus often termed "static" or "structural". If they depend only on the positions they can also be obtained from Monte-Carlo (MC) simulations, in which the ensemble is constructed through a random-walk procedure.

6.2. Time-Correlation Functions

The time-correlation function $c_{BC}(t)$ between two quantities B and C is defined in the equilibrium ensemble as Equation (18)

$$c_{BC}(t) = \langle B(\mathcal{R}(\tau)) \otimes C(\mathcal{R}(t+\tau)) \rangle_\tau \tag{18}$$

where $B(\mathcal{R})$ and $C(\mathcal{R})$ are quantities like $A(\mathcal{R})$ above, $\otimes$ denotes a general operator (most often a product or scalar product, sometimes also a logical AND or OR operator) and $\langle\ \rangle_\tau$ denotes an average over all times τ. If $C = B$, $c_{BB}(t)$ is called an autocorrelation function. Otherwise, it is called a cross-correlation function.

Example 3. One of the simplest and most widely studied autocorrelation functions is the velocity autocorrelation function, i.e., $B(\mathcal{R}) = \vec{V}$, where $\vec{V}$ is the vector of the velocities of N equivalent particles (e.g., the same type of atom, or a given type of atom fulfilling some condition) at a given time (Eq. 19).

$$\vec{V}(\mathcal{R}(t)) = \begin{pmatrix} \vec{v}_1 \\ \vec{v}_2 \\ \vdots \\ \vec{v}_N \end{pmatrix} \tag{19}$$

The velocity autocorrelation function then becomes in terms of the $\vec{v}_i$ Equation (20)

$$\begin{aligned} c_{\vec{V}\vec{V}}(t) &= \left\langle \vec{V}(\mathcal{R}(\tau)) \cdot \vec{V}(\mathcal{R}(t+\tau)) \right\rangle_\tau \\ &= \frac{1}{N \cdot M} \sum_{j=0}^{M} \sum_{i=1}^{N} \left(\vec{v}_i(\tau_j) \cdot \vec{v}_i(\tau_j + t) \right) \\ &= \left\langle c_{\vec{v}_i \vec{v}_i}(t) \right\rangle \end{aligned} \tag{20}$$

Equation (20) can be easily evaluated from the trajectory; an example is given in Sect. Examples

Integrals of correlation functions of the type Equation (21)

$$\mathrm{T} = \lim_{t \to \infty} \int_0^t c_{AA}(\tau) \mathrm{d}\tau \tag{21}$$

are related to transport properties, in the case of the velocity autocorrelation function, to the self-diffusion coefficient D_S. For the integral to converge, the correlation function must decay within the simulation time. Actually, it must decay much faster so that a sufficient number of averages (the sum over j in Eq. 20) can be carried out. If the expectation value of the autocorrelated quantity is not equal to zero, the definition of the function is usually modified so that it indeed decays to zero at long times.

Classical time correlations are even in time. It is thus sufficient to consider their Fourier–Laplace transforms. For a function $c_{AA}(\tau)$, the transform (Eq. 22)

$$\tilde{c}_{AA}(\omega) \propto \int_0^\infty c_{AA}(\tau) \cos(\omega\tau) \mathrm{d}\tau \tag{22}$$

is called the spectral density, or the simply the spectrum. Depending on which quantity A is correlated, the spectra can be observed experimentally (e.g. in IR spectroscopy, corresponding to the correlation of the dipole moment or derivative of the dipole moment of the system). The velocity autocorrelation function is again used as an example in Chapter Examples.

[Top of Page]

7. Examples

The first example is a simulation of a system consisting of 864 Lennard–Jones particles (Eq. 7) parameterized to represent methane (CH_4, united-atom approximation, effective potential from Figure 2) in an NVE simulation with volume and energy chosen to yield an average temperature (Eq. 17) of $\langle T \rangle$ = 130 K and a density of 0.3952 g/cm^3. Figure 6 shows the simulated radial pair distribution function (RDF) $g(r)$, which is a measure of the microscopic structure in a liquid [29]. It represents the relative density, compared to the overall bulk density, as a function of distance from a fixed particle at $r = 0$. In the region of the first peak, located roughly where the Lennard–Jones potential has its minimum, the local density is thus enhanced by a factor of approximately 2.6. These density fluctuations extend to further "shells" of neighbors, and finally the $g(r)$ function converges toward a value of 1 (bulk density). RDFs can be determined experimentally, under certain approximations, from X-ray, neutron, or other scattering methods; they are also an important tool for the computation of thermodynamics quantities in statistical mechanics.

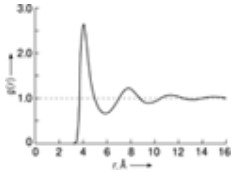

Figure 6. Radial distribution function $g(r)$ of liquid methane at $\langle T \rangle$ = 130 K and a density of 0.3952 g/cm^3.

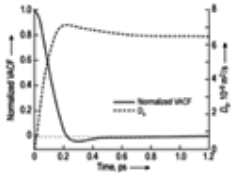

Figure 7. Normalized VACF of liquid methane and integral of the unnormalized function, yielding at long times a self-diffusion coefficient of $D_S \approx 6.66 \times 10^{-9}$ m^2/s.

As outlined in Section Time-Correlation Functions, dynamical quantities are accessible by computing suitable time-correlation functions. The self-diffusion coefficient D_S of liquid methane is thus obtained from the molecular velocity autocorrelation function (VACF) using Equations (20) and (21). The underlying simulation was run under the condition of zero total momentum $\vec{p}$ = 0 (Eq. 15), i.e., the expectation value of a particle velocity is zero (see above). Figure 7, left axis, shows this VACF, normalized to unity at t = 0. The right-hand scale in Figure 7 refers to an integral like in Equation (21), suitably normalized to yield at long times the self-diffusion coefficient D_S.

To illustrate the real movements of some methane molecules during a simulation, Figure 8 displays the projections onto a plane of four particle trajectories generated during 25 ps of simulation time. We note that this time is much longer than the correlation time of the velocities in Figure 7. Such plots are useful to gauge how extensively the thermodynamic ensemble has been sampled by the simulation.

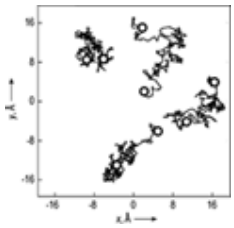

Figure 8. Trajectories of four different arbitrarily chosen methane molecules over 25 ps of simulation. The frame represents the quadratic simulation box with a side length of 38.8 Å. Initial and end points are marked by a bold circle.

[Top of Page]

8. Applications

The applications of MD simulations are by now almost innumerable, and only a few examples can be given here. The first examples, discussed in the previous sections, were taken from studies of liquids such as liquid methane or aqueous ionic solutions at ambient conditions, for which much of the basic methodology was developed. However, one of the main advantages of simulation methods is that no assumptions have to be made concerning the state of the system: gases, liquids, and solids, as well as inhomogeneous systems containing all kinds of interfaces, are thus, in principle, amenable to MD simulations. As a caveat, we note, however, that in some cases this method will be quite inefficient. In gases at low densities, for instance, very much time is spent integrating the motions of molecules with almost negligible mutual interactions.

It is easy in MD simulations to study systems under extreme conditions of density (pressure) and temperature; such systems are usually difficult to study experimentally. Thus, molten salts were among the first systems studied by simulation techniques. Studies of liquids of geological interest, i.e., usually at pressures and temperatures difficult to maintain in a laboratory experiment, are other examples for this kind of approach. Supercritical fluids in thermodynamic conditions not too close to the critical point can be investigated in the same fashion. Studies on highly complex liquids such as polymer melts or liquid crystalline systems are becoming more and more numerous as the available computational resources increase and models and methods for their simulations are developed.

Besides neat liquids and solutions, the structure and dynamics of liquid mixtures has become a very active field of investigation. We note here, however, that studies of certain thermodynamic properties, in particular of the phase equilibria, are most often undertaken by using specific Monte-Carlo (MC) techniques rather than MD simulations. The development of adequate interaction models, i.e., ones able to represent both the neat liquids and the mixtures, is one of the main challenges when mixtures are to be studied.

Among the many applications concerning interfaces (liquid–liquid, liquid–gas, and solid–liquid) the free surfaces of pure liquids and solutions have been studied. Furthermore, liquids near smooth hydrophilic or hydrophobic walls with or without corrugation have been extensively simulated. Systems of electrochemical interest, e.g., solid metal electrodes in contact with aqueous ionic solutions, but also confined liquids such as liquids in pores or in clays, have been of particular interest. The dissolution and crystallization processes of salts have been investigated in simulations of salt solutions in contact with the corresponding crystals. Layers of amphiphilic molecules, soap films, micelles, and micellar solutions of various kinds

have been investigated in very extensive MD simulations, as have polymer layers and model biomembranes. In some of the studies on interfacial or porous systems it is sufficient to consider the solid phase as an immobile “frozen” obstacle. However, if the dynamics of such crystalline solids are to be simulated, which is often necessary, the restrictions imposed by the periodic boundary conditions must be considered. Multiples of the elementary cell are used as simulation boxes, and special techniques are available to adjust the box, e.g., if structural phase transitions are to be studied. Multiple studies on inclusion compounds such as clathrate hydrates and zeolites, with or without guest molecules, have been carried out in this way.

The biosciences are by now possibly the largest application domain for MD simulations. The main problems here are the sizes of the systems and the consequent complexity of the interaction models. Furthermore, in many instances these systems are characterized by correlation times that are long compared to the times usually accessible in MD simulations (see Figure 5). Extremely long simulation runs are thus necessary, and present-day resources often do not allow a satisfactory sampling. Yet events like the penetration of small solvated molecules, or indeed water itself, into membranes or trans-membrane ion channels have been studied by simulation techniques. Derived techniques are used in drug design. Understanding protein folding remains one of the challenges in this area, and large-scale efforts are underway. This is, however, beyond the scope of this article.

As discussed above, transport coefficients can be obtained, within the framework of the fluctuation-dissipation theorem [6], from correlation functions computed for the system in equilibrium (see Section Time-Correlation Functions). However, it is often more efficient to perform nonequilibrium molecular dynamics (NEMD). This can be done both for stationary (e.g., in a thermal gradient) and nonstationary (e.g. for relaxation phenomena) conditions. Many methods have been proposed [12], [30-32]. Essentially, in the first instance, gradients are maintained across the system during the simulation in a way compatible with the periodic boundary conditions. In the second instance, an equilibrium simulation is performed and configurations are periodically extracted, perturbed, and their evolution is compared with the ongoing equilibrium run. The NEMD simulation allows one to go beyond the linear response and to study nonlinear transport phenomena.

In a coarse-graining approach, macroscopic quantities obtained from MD simulations are often used as input parameters to other types of simulations, e.g. Brownian Dynamics [12] or dissipative particle dynamics (DPD) [33]. In these approximations, much larger systems can be studied over longer periods of time, going well beyond the domain sketched in Figure 5.

Finally, we note that more and more approaches combining MD simulations (in thermodynamic equilibrium, as discussed here, or out of equilibrium) with other approaches are being proposed for various purposes. Chemical reactions are thus tackled by combining MD with quantum chemistry calculations. Other efforts aim at extending the space and time domains accessible to the simulations by using suitable coarse-graining techniques. These methods are mostly still in the stage of methodic developments. Like the now classical methods discussed here, which were developed in the last third of the 20th century, they will one day become tools for research in all kinds of disciplines.

[Top of Page]

References

General textbooks

M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids*, Clarendon Press, Oxford, 1990.

K. Binder, *The Monte Carlo Method in Condensed Matter Physics*, Springer-Verlag, Berlin, 1992.

G. Ciccotti, D. Frenkel, I. R. McDonald, *Simulation of Liquids and Solids: Molecular Dynamics and Monte Carlo Methods in Statistical Mechanics*, North-Holland, Amsterdam, 1987.

D. J. Evans, G. P. Morriss, *Statistical Mechanics of Non-Equilibrium Liquids*, Academic Press, London, 1990.

D. Frenkel, B. Smit, *Understanding Molecular Simulation — From Algorithms to Application*, Academic Press, San Diego, 2002.

K. E. Gubbins, N. Quirke (Eds), *Molecular Simulation and Industrial Applications. Methods, Examples and Properties*, Gordon and Breach, Amsterdam, 1996.

R. Haberlandt, S. Fritzsche, G. Peinel, K. Heinzinger, *Molekulardynamik — Grundlagen und Anwendungen*, Vieweg Verlag, Braunschweig, 1995.

J. M. Haile, *Molecular Dynamics Simulation: Elementary Methods*, Wiley, New York, 1992.

D. M. Heyes, *The Liquid State: Applications of Molecular Simulation*, Wiley, Chichester, 1998.

W. G. Hoover, *Computational Statistical Mechanics*, Elsevier, Amsterdam, 1991.

G. C. Maitland, M. Rigby, E. B. Smith, W. A. Wakeham, *Intermolecular Forces: Their Origin and Determination*, Clarendon Press, Oxford, 1981.

D. C. Rapaport, *The Art of Molecular Dynamics Simulation*, Cambridge University Press, Cambridge, 1995.

R. J. Sadus, *Molecular Simulation of Fluids — Theory, Algorithms and Object-Orientation*, Elsevier, Amsterdam, 1999.

M. Schoen, *Computer Simulation of Condensed Phases in Complex Geometries*, Springer-Verlag, Berlin, 1993.

F. Vesely, *Computerexperimente an Flüssigkeitsmodellen*, Physik Verlag, Weinheim, Germany, 1978.

F. Vesely, *Computational Physics. An Introduction*, Kluwer Academic/Plenum Publishers, New York - London, 2001.

Selected Articles

K. Binder, J. Horbach, W. Kob, W. Paul, F. Varnik, “Molecular Dynamics Simulations”, *J. Phys. Condens. Matter* **16** (2004) 429–453.

G. Ciccotti, G. Jacucci, I. R. McDonald, “Thought-Experiments by Molecular Dynamics”, *J. Stat. Phys.* **21** (1979) no. 1, 1–22.

W. G. Hoover, W. T. Ashurst, “Nonequilibrium Molecular Dynamics” in *Theoretical Chemistry. Advances and Perspectives*, **Vol. 1**, Academic Press, San Diego, 1975, pp. 2–51.

W. Smith, C. W. Young, P. M. Rodger, “DL_POLY: Application to Molecular Simulation”, *Mol. Sim.* **28** (1975) no. 5, 385–471.

W. F. van Gunsteren, H. J. C. Berendsen, “Computer Simulation of Molecular Dynamics: Methodology, Applications and Perspectives in Chemistry”, *Angew. Chem. Int. Ed. Engl.* **29** (1990) 992–1023.

Web resources (as of 2004):

A glossary of terms: http://xeon.concord.org:8080/modeler1.2/guide/html/glossary/index.html

A. S. Côté, B. Smith, P. J. D. Lindau, *Democritus — A Molecular Dynamics Tutorial*: http://www.compsoc.man.ac.uk/~lucky/Democritus/Basic/Democritus.html, Daresbury Laboratory, Daresbury, UK, 2001.

P. Cummings, D. A. Kofke, A. Panagiotopolous, J. de Pablo, R. Rowley, H. Cochran, D. Evans, *Molecular Simulation. A World Wide Web Book*, Written by a task force and published under the auspices of the National Science Foundation: http://boltzmann.vuse.vanderbilt.edu/~w3press/.

F. Ercolessi, A Molecular Dynamics Primer: http://www.fisica.uniud.it/~ercolessi/md/md/, June 1997. University of Udine, Italy.

D. A. Kofke, Molecular Simulation Lectures: http://www.cheme.buffalo.edu/courses/ce530/index.html, Department of Chemical Engineering, SUNY Buffalo.

F. Vesely. Computational Physics. Course material: http://www.ap.univie.ac.at/users/ves/cp0102/dx/index.html, Institute of Experimental Physics, University of Vienna, 2002.

F. Vesely. Statistical Physics. Course material with JAVA applets: http://www.ap.univie.ac.at/users/ves/sp_english/sp/sp.html, Institute of Experimental Physics, University of Vienna, 2002.

Specific references

1. K. Binder, D. W. Heermann, *Monte Carlo Simulation in Statistical Physics*, Springer-Verlag, Berlin, 1997.
2. The Gallup Organization, http://www.gallup.com, Princeton, NJ, accessed 2004.
3. P. W. Atkins, J. de Paula, *Atkins' Physical Chemistry*, Oxford University Press, Oxford, 2002.
4. J. B. Foresman, A. Frisch, *Exploring Chemistry with Electronic Structure Methods*, Gaussian, Inc., Pittsburgh, PA, 1993.
5. A. Hinchliffe, *Computational Quantum Chemistry*, Wiley, New York, 1988.
6. D. A. McQuarrie, *Statistical Mechanics*, Harper & Row, New York, London, 1976.
7. R. Car, M. Parrinello, “Unified Approach for Molecular Dynamics and Density-Functional Theory”, *Phys. Rev. Lett.* **55** (1985) 2471–2474. Links
8. J. M. Haile, *Molecular Dynamics Simulation: Elementary Methods*, Wiley, New York, 1992.
9. B. Guillot, “A Reappraisal of What We Have Learnt During Three Decades of Computer Simulations on Water”, *J. Mol. Liq.* **101** (2002) nos. 1–3, 219–260. Links
10. E. B. Wilson, J. C. Decius, P. C. Cross, *Molecular Vibrations*, McGraw-Hill, New York, 1955.
11. G. Herzberg, *Molecular Spectra and Molecular Structure I*, Van Nostrand, New York, 1950.
12. M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids*, Clarendon Press, Oxford, 1990.
13. C. L. Kong, “Combining Rules for Intermolecular Potential Parameters. II. Rules for the Lennard-Jones (12-6) Potential and the Morse Potential”, *J. Chem. Phys.* **59** (1973) 2464. Links
14. J. O. Hirschfelder, C. F. Curtiss, R. B. Bird, *Molecular Theory of Gases and Liquids*, Wiley, New York, 1954.
15. J. G. Gay, B. J. Berne, “Modification of the Overlap Potential to Mimic a Linear Site-Site Potential”, *J. Chem. Phys.* **74** (1981) 3316. Links
16. G. C. Maitland, M. Rigby, E. B. Smith, W. A. Wakeham, *Intermolecular Forces: Their Origin and Determination*, Clarendon Press, Oxford, 1981.
17. B. M. Axilrod, E. Teller, “Interactions of the van der Waals' Type between Three Atoms”, *J. Chem. Phys.* **11** (1943) 299–300. Links
18. A. P. Sutton, J. Chen, “Long-Range Finnis-Sinclair Potentials”, *Philos. Mag. Lett.* **61** (1990) no. 3, 139–146. Links
19. J. D. Jackson, *Classical Electrodynamics,* Wiley, New York, 1975.
20. L. X. Dang, “A Mechanism for Ion Transport across the Water/Dichloromethane Interface: A Molecular Dynamics Study Using Polarizable Potential Models”, *J. Phys. Chem. B* **105** (2001) 804–809. Links
21. J.-C. Soetens, G. Jansen, C. Millot, “Molecular Dynamics Simulation of Liquid CCl_4 with a New Polarizable Model”, *Mol. Phys.* **96** (1999) 1003. Links
22. H. Meyer, O. Biermann, R. Faller, D. Reith, F. Müller-Plathe, “Coarse Graining of Nonbonded Inter-particle Potentials Using Automatic Simplex Optimization to Fit Structural Properties”, *J. Chem. Phys.* **113** (2000) 6264–6275. Links
23. W. G. Hoover, *Computational Statistical Mechanics*, Elsevier, Amsterdam, 1991.
24. S. Toxvaerd, O. H. Olsen, “Canonical Molecular Dynamics of Molecules with Internal Degrees of Freedom”, *Ber. Bunsenges. Phys. Chem.* **94** (1990) 274. Links
25. H. Flyvberg, H. G. Petersen, “Error Estimates on Averages of Correlated Data”, *J. Chem. Phys.* **91** (1989) 461. Links
26. R. Lustig, “Statistical Thermodynamics in the Classical Molecular Dynamics Ensemble. I. Fundamentals”, *J. Chem. Phys.* **100** (1994) 3048–3059. Links
27. R. Lustig, “Statistical Thermodynamics in the Classical Molecular Dynamics Ensemble. II. Application to Computer Simulation”, *J. Chem. Phys.* **100** (1994) 3060–3067. Links

28. R. Lustig, “Statistical Thermodynamics in the Classical Molecular Dynamics Ensemble. III. Numerical Results”, *J. Chem. Phys.* **100** (1994) 3068–3078. Links
29. J. P. Hansen, I. R. McDonald, *Theory of Simple Liquids*, Academic Press, New York, 1986.
30. D. J. Evans, G. P. Morriss, *Statistical Mechanics of Non-Equilibrium Liquids*, Academic Press, London, 1990.
31. W. G. Hoover, “Nonequilibrium Molecular Dynamics”, *Ann. Rev. Phys. Chem.* **34** (1983) 103–127. Links
32. W. G. Hoover, W.T. Ashurst, “Nonequilibrium Molecular Dynamics” in *Theoretical Chemistry. Advances and Perspectives, Vol. 1*, Academic Press, San Diego, 1975, pp. 2–51.
33. D. Frenkel, B. Smit, *Understanding Molecular Simulation — From Algorithms to Application*, Academic Press, San Diego, 2002.

[Top of Page]

Symbols and Abbreviations

***A*, *B*, *C*:** state functions, i.e., functions that can be computed knowing a configuration

A_{ij}, B_{ij}**:** parameters of the Lennard–Jones (12–6) potential

$\overleftrightarrow{\alpha}$**:** molecular polarizability tensor

$c_{BC}(t)$**:** time correlation function between state functions B and C

***d*:** interparticle distance

***D*:** depth of Morse potential

D_s**:** self-diffusion coefficient

δt**:** time step size

|*e*|: elementary electric charge

$\vec{E}$**:** electric field

***E*:** total energy of the system

E^{kin}**:** kinetic energy

ε**:** parameter of the Lennard–Jones (12–6) potential and of potentials derived from the LJ potential

$\vec{F}_i$**:** force acting on particle *i*

Υ**:** constant of the Axilrod–Teller potential

***h*:** Planck's constant, $h = 6.62608 \times 10^{-34}$ J s

$\vec{k}$**:** wave vector

k_B**:** Boltzmann's constant, $k_B = 1.38066\ 10^{-23}$ J/K

k_{lm}, k_{lmn}, k_{lmno}**:** (generalized) force constants

Λ**:** De Broglie wavelength

m_i**:** mass of particle *i*

***M*:** number of events or occurrences

M_{dof}**:** degrees of freedom

***N*:** total number of particles in a simulated system

ω**:** angular frequency

***p*:** pressure

$\vec{p}$**:** momentum

$\vec{P}$**:** molecular polarization

q: electric (partial) charge, i.e. multiples of the elementary charge |*e*|

$\mathscr{R}$: configuration

$\vec{r}_i$: vector x_i, y_i, z_i of particle coordinates

$\ddot{\vec{r}}_i$: acceleration of particle *i*

$\vec{r}_{ij}$: position of site *i* relative to *j*, $(\vec{r}_i - \vec{r}_j)$

r_{cut}: cut-off radius

ρ: internal coordinate

S: length of the periodic box

σ: parameter of the Lennard–Jones (12–6) potential

t, τ: time

T: absolute temperature

T: transport property

θ: angle

U: potential energy

V: volume

$\vec{V}$: vector of velocities

$\vec{v}_i$: vector V_{x_i}, V_{y_i}, V_{z_i} of particle velocities, $\vec{v}_i = \dot{\vec{r}}_i$

$\langle\ \rangle$: average over configurations

CPMD: Car–Parrinello molecular dynamics

DFT: density functional theory

DPD: dissipative particle dynamics

LJ: Lennard–Jones

MD: molecular dynamics

MC: Monte Carlo

NpT: isothermal–isobaric ensemble

NVE: microcanonical ensemble

NVT: canonical ensemble

NEMD: nonequilibrium molecular dynamics

RDF: radial distribution function *g*(*r*)

VACF: velocity autocorrelation function

[Top of Page]

Related Articles

See also ⟶ Molecular Modeling.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Membranes: Separation of Organic Vapors from Gas Streams

Klaus Ohlrogge, GKSS Forschungszentrum Geesthacht GmbH, Geesthacht, Federal Republic of Germany
Kai Stürken, Sterling Sihi GmbH, Itzehoe, Federal Republic of Germany

1. Introduction 1
2. Membranes 1
3. Membrane Selectivity 2
3.1. Temperature, Pressure, and Feed-Gas Concentration 3
3.2. Real-Gas Behavior 3
3.3. Concentration Polarization 4
3.4. Joule – Thomson Effect 4
4. Applications 5
4.1. Gasoline Vapor Recovery 5
4.2. Retrofitting of Activated Carbon Based Vapor Recovery Units 6
4.3. Emission Reduction at Gas Stations . 7
4.4. Off-Gas Treatment in the Chemical Industry 8
4.5. Recovery of Vinyl Chloride Monomer from Off-Gas of PVC Production . . . 9
4.6. Batch Reactions and Vacuum Drying of Fine Chemicals 10
4.7. Polyolefin Production Processes 11
4.8. Oil and Gas Production 13
5. References 15

1. Introduction

Many effluents emitted from chemical, petrochemical, and pharmaceutical production plants contain volatile organic compounds (VOCs). Furthermore, hydrocarbon vapors are generated by handling, storage, and distribution of solvents and gasoline. Hydrocarbon vapors are known as ozone precursors and are one facsstor causing the formation of summer smog. In addition, hydrocarbon vapors are believed to be greenhouse gases. Aromatic hydrocarbons such as benzene are considered to be carcinogenic. The introduction of new clean air acts with new and more stringent emission standards has generated a driving force to develop new and more efficient off-gas treatment units. Membrane-based separation technologies for the recovery of organic vapors from gaseous effluents were introduced on a commercial scale in the early 1990s [1]. Since ca. 1995 the acceptance of this technology has been growing because of its confirmed performance, reliability, and economic operation. Based on the experience gained with end-of-pipe applications, the use of membranes in integrated processes to separate organic compounds has been explored. Well-accepted applications are the use of membranes in polymer production processes for separation of olefins (ethylene, propylene) from process streams, separation of hexane and vinyl chloride monomer, and nitrogen purification. Newly explored applications are hydrocarbon dewpointing of natural gas, treatment of associated gas from oil production, and control of methane number of the feed gas for combustion engines.

2. Membranes

Membranes which are commonly used in organic vapor separation applications are flat-sheet thin-film composite membranes. Typically, the membrane consists of a multilayer structure. Nonwoven materials, e.g., polyester or polypropylene fleeces, are used as substructure to provide the necessary mechanical strength. A microporous substrate made from polysulfone, polyimide, polyacrylonitrile, or poly(vinylidene fluoride) is cast on the fleece [2]. Pore size and pore distribution of the substructure must be formed in accordance with the viscosity and adherence of the coating solution which is applied on top of the microporous substrate to form the dense selective film. Common membrane polymers to separate organic vapors from permanent gases are polydimethylsiloxane (PDMS) and polyoctylmethylsiloxane (POMS) [3]. Dense

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.j16_j01

thin films with a thickness of 1 – 5 µm have adequate selectivity and sufficient permeance for a variety of organic vapors. Compared to PDMS, POMS has higher selectivity but lower permeability. The advantage of the higher selectivity is the reduced copermeation of compounds to be retained on the feed side of the membrane. This results in reduced sizes of feed and permeate pumps but also in increased membrane area. Investment and operating costs must be considered when selecting the membrane polymer. The flat-sheet membranes are employed in spiral-wound or envelope-type modules for technical configurations.

3. Membrane Selectivity

Gas transport through a dense film depends on the interactions of the feed-gas compounds with the membrane polymer regarding its solubility and diffusivity. The diffusion coefficient of a gas molecule is expected to decrease with increasing molecular size. This is true in the case of glassy polymers with stiff polymer backbones, which act like molecular sieves. Small molecules like helium, hydrogen, or nitrogen permeate much faster through the rigid polymer backbone than larger molecules such as butane or pentane. An elastomeric polymer behaves more like a liquid. Gas transport through rubbery polymers is more determined by its solubility than its diffusivity. The diffusion and solubility coefficients of different gases in silicone rubber are depicted in Figure 1 [4].

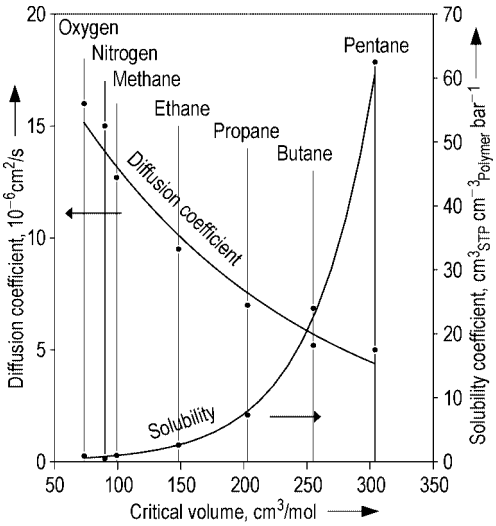

Figure 1. Diffusion and solubility coefficients of different gases in silicone rubber at 30°C [5 – 7] versus the critical volume

It is apparent that the slower diffusion of hydrocarbons like propane, butane, and pentane is overcompensated by their higher solubility. As a result silicone rubber is more permeable for hydrocarbons than for permanent gases like oxygen or nitrogen. The plot in Figure 2 shows the selectivity of various hydrocarbon vapors over nitrogen. The organic compounds are arranged with respect to their material category.

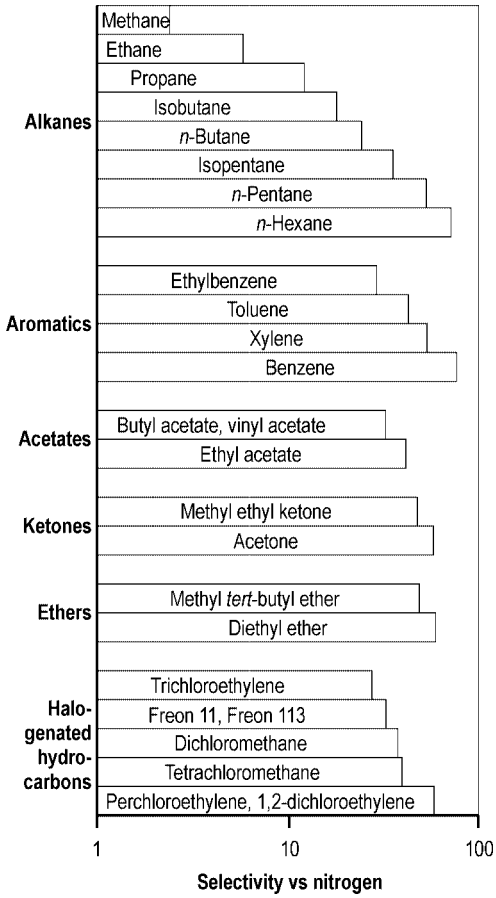

Figure 2. Selectivities of various organic vapors versus nitrogen

The length of the bars shows the highest selectivity determined by single-gas measurements at ambient temperature. The selectivities obtained in real technical membrane separation processes are lower and depend on process

and design parameters like pressure, temperature, feed-gas composition, gas velocity over the membrane on the feed side, pressure drop in the membrane module, and module and membrane design. Especially the impact of concentration polarization caused by the formation of a boundary layer adjacent to the membrane surface has a strong impact on the overall selectivity of the separation process.

3.1. Temperature, Pressure, and Feed-Gas Concentration

The process operating temperature has a direct influence on the membrane selectivity. The diffusion coefficient of gas molecules increases with increasing temperature. However, the permeance of organic vapors through elastomeric polymers is governed by their solubility, which increases with decreasing temperature. Hence, the permeation of organic compounds increases with decreasing temperatures, whereas the permeation of diffusion-driven compounds like nitrogen or oxygen decreases (Fig. 3). It is advantageous to perform separation of organic vapor at the lowest possible operating temperature in order to achieve the highest possible membrane selectivity.

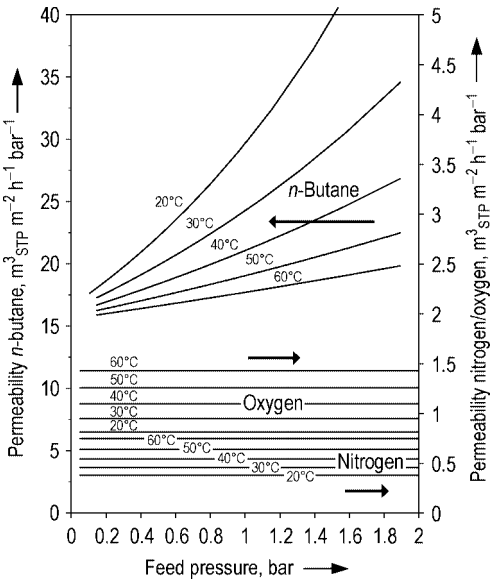

Figure 3. Temperature dependence of gas and vapor permeation through PDMS membranes

Silicone-based membrane materials swell in the presence of organic compounds depending on their vapor pressure. The presence of permanent gases like oxygen, nitrogen, or hydrogen has no impact on membrane swelling. The sorption of organic compounds plasticizes the membrane material and increases the diffusion coefficient. Membrane swelling leads to a higher permeance of all gas and vapor compounds (Fig. 4).

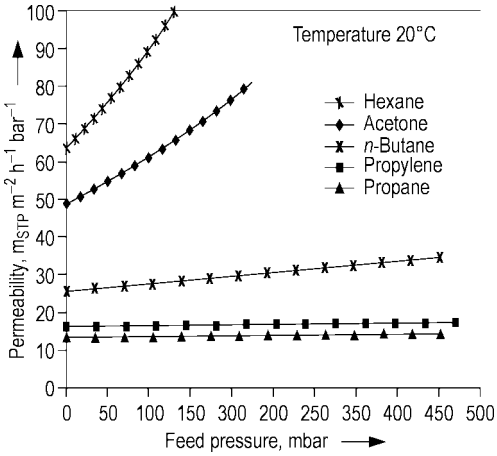

Figure 4. Flux densities of organic vapors and nitrogen over vapor pressure (single-gas measurements)

3.2. Real-Gas Behavior

The membrane permeance is defined by

$$L_i = \frac{\dot{V}_{i\,\mathrm{P(STP)}}}{A_\mathrm{M} \cdot (p_{i\mathrm{H}} - p_{i\mathrm{L}})} \tag{1}$$

where L_i is the permeance, $\dot{V}_{i\,\mathrm{P(STP)}}$ is the permeate volume rate in the standard state, A_M is the membrane area, and $p_{i\mathrm{H}}$ and $p_{i\mathrm{L}}$ are the partial pressures of components at the feed and permeate side of the membrane, respectively. The partial pressures must be replaced by fugacities in medium- and high-pressure applications in order to avoid a faulty process design.

At pressures exceeding 10 bar the driving force of higher hydrocarbons is reduced by the real-gas behavior. The difference in partial pressures must be replaced by the difference in fugacities. A comparison of the partial pressure difference and fugacity difference driving forces versus total feed pressure is plotted in Figure

5 for a binary mixture with 5 % n-butane in methane.

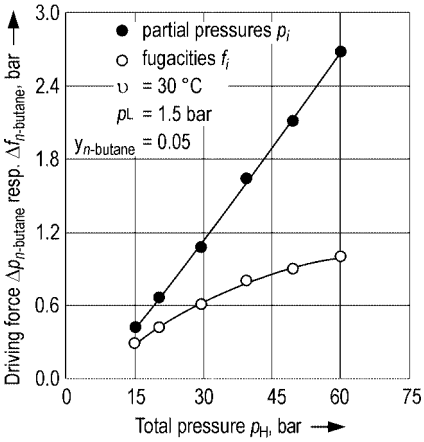

Figure 5. Calculated ideal and real driving force of n-butane in a binary mixture of n-butane and methane [8]

The dependence of fugacities of some organic vapors on pressure is depicted in Figure 6. Clearly, real-gas behavior cannot be neglected at operating pressure exceeding 10 bar.

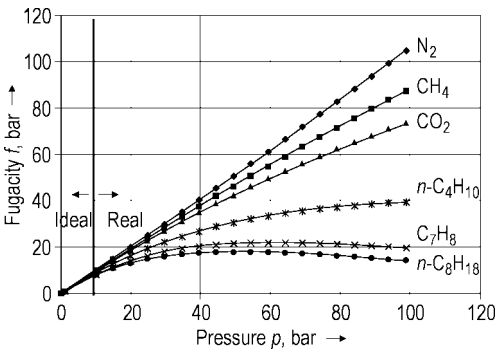

Figure 6. Influence of pressure on real gas behavior of organic vapor compounds [9]

3.3. Concentration Polarization

The design and selection of membrane and membrane module is highly dependent on process and operating conditions. The thickness of the selective layer of the composite membrane should be in accordance with permeance and selectivity of feed compounds, as well as the Reynolds number of the feed volume flow over the membrane surface. The feed flow Reynolds number has a direct influence on the thickness of the boundary layer between the bulk gas and the membrane surface. The thickness of the selective layer has a reciprocal dependency on the gas and vapor permeance. The porous substructure of the membrane must be as open as possible to provide a nearly unrestrained flux of all permeating compounds. Boundary layer, selective layer, and porous support create the major transport resistances (Fig. 7). The behavior of the selective layer and the buildup of the boundary layer are interdependent. A high permeance of a fast compound will increase the influence of the boundary layer, resulting in high concentrations of the slow compounds and low concentrations of the fast compounds on the membrane surface. This effect decreases the driving force for the preferentially permeating component (cf. Eq. 1). Transport through the boundary is of diffusive nature. Hence, concentration polarization is of increasing importance with increasing operating pressure, since the diffusion coefficient is inversely proportional to pressure.

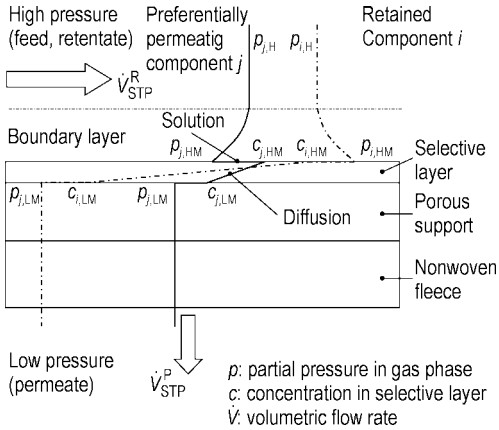

Figure 7. Gas transport through polymer membranes

3.4. Joule – Thomson Effect

The Joule – Thomson effect must be taken in account if the transmembrane pressure difference is high [10]. The temperature change in a membrane module is caused by two effects:

- Adiabatic throttling of the permeating gas

- Heat transfer from the high- to the low-pressure side of the membrane

The Joule – Thomson effect is of major importance for high-stage-cut separation. The stage cut is defined as permeate volume flow over feed volume flow. In this case it is insufficient to design a gas separation process solely in terms of material balances, permeation equations, and mass transfer kinetics. The energy balances must also be taken in account. The temperature change along the membrane module has a direct influence on permeances and selectivities, since they are functions of temperature. In case of natural gas applications a heat-recovery system in front of the separation stage could improve the overall process efficiency [11].

4. Applications

4.1. Gasoline Vapor Recovery

Fuel distribution is considered to be an area with a major potential for reduction of VOC emissions. The principle of emission reduction in the chain of gasoline distribution from refinery to tank farm, gas station, and car refueling is vapor balancing. The volume of vapor which is displaced when a volume of liquid is filled into a tank is collected and returned to the tank from which the liquid is drawn off. The gas volume is increased because of possible changes of temperatures and pressures as well as by evaporation of low-boiling substances, which are generated by turbulences of the liquid phase during filling procedures. If the principle of vapor collection is consistently applied throughout the complete chain of gasoline distribution, nearly all generated vapors can be separated and recovered in a central vapor recovery unit (VRU) at a refinery or tank farm. The average gasoline vapor concentration in off-gases from tank farms is approximately 20 – 40 vol %, which corresponds to approximately 600 – 1200 g hydrocarbons/m^3 air. This amount also represents a commercial value which should be recovered, especially if the distributed gasoline was taxed. The stringent air-quality control requirements are laid down in clean air acts, ordinances on the reduction of hydrocarbon emissions, or technical instructions on air quality control [12 – 14]. Most of the provisions are considerably more stringent than the European Directive 94/63/EC, which allows an emission limit value for hydrocarbons of 35 g/m^3. This is far from reflecting best available technology (BAT) of vapor recovery units. It is also dangerously close to the lower explosion limit (40 g/m^3) of gasoline vapors and thus represents a safety risk.

Gasoline vapors are generated by tank filling, storage-tank ventilation, and operation of truck filling stations, rail wagon filling stations, or ship loading facilities. If a constant off-gas flow is expected, the off-gas can be directly fed to the vapor recovery unit. In case of fluctuating off-gas volumes it can be recommended to install a gas holding tank to collect vapors between the emission source and the vapor recovery unit in order to smooth out volume peaks. Buffering of off-gas volumes leads to a plant design according to an average off-gas flow and a continuous plant operation (Fig. 8).

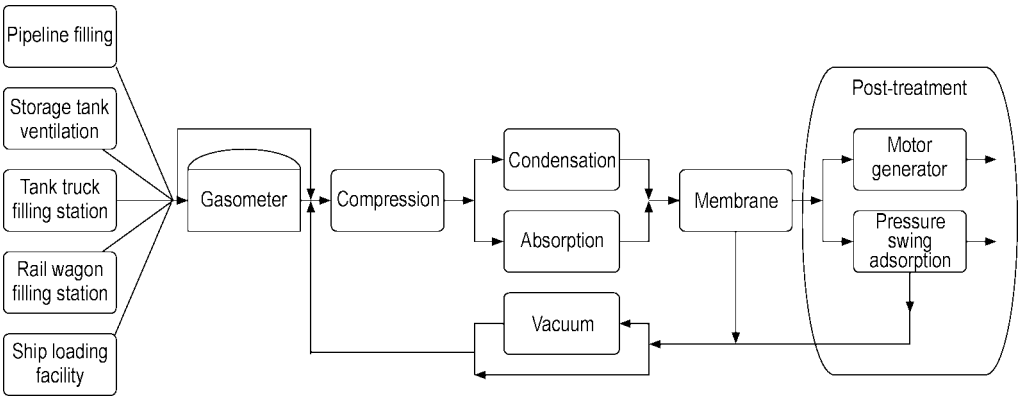

Figure 8. Flow sheet for gasoline vapor recovery technology

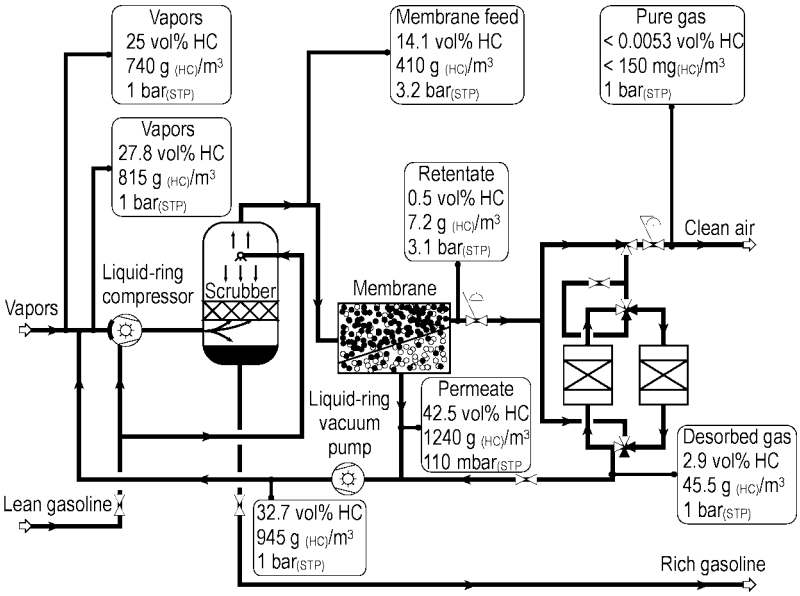

Figure 9. Flow scheme of a hybrid vapor recovery unit consisting of a membrane stage and pressure swing adsorption unit

The first processing step is pressure increase. Typically, liquid-ring compressors or screw compressors are used. The gasoline is recovered by condensation or absorption before the off-gases are fed to a membrane stage. The membrane stage is designed according to the required vent-gas purity. The membrane separates hydrocarbon vapors from air. Hydrocarbon permeates through the membrane, whereas the air remains on the feed side of the membrane. A pressure ratio (feed pressure over permeate pressure) of 6 – 10 must be applied to support the membrane separation process. Liquid-ring compressors operate at 3 – 3.5 bar, whereas the screw compressors operate at a feed pressure of approximately 6 – 9 bar.

In liquid-ring compressor applications, the membrane separation process is supported by applying vacuum at the permeate side. The permeate stream, enriched in hydrocarbon vapor, is recycled to the intake of the feed compressor. Single-stage membrane units are very efficient up to a residual vent-gas concentration of 4–35 g/m^3 of organic compounds. If stringent clean air acts require lower outlet concentrations post-treatment by pressure swing adsorption or combustion can be recommended. A hybrid system consisting of a membrane stage and a polishing step by pressure swing adsorption (Fig. 9) is the best-selling system combination.

This combination has some essential advantages. The operating pressure of the membrane process can be used advantageously to support the adsorption stage, whereas the vacuum can be used to support the desorption of the regenerating adsorber column. Because of the low and defined retentate concentration the temperature increase caused by adsorption is less than 5°C. The system is also characterized by the ease of shutdown and startup (Fig. 10).

4.2. Retrofitting of Activated Carbon Based Vapor Recovery Units

Existing activated carbon based vapor recovery units erected in the early 1980s often do not meet current clean air standards, which are adapted to developments in separation techniques. A possibility to improve the overall performance of an adsorption unit is the installation of a membrane stage to precondition the feed stream to the adsorption unit. By means of the membrane stage higher hydrocarbons and peak concentrations are separated and a stream nearly constant in hydrocarbon concentration is fed to the adsorption plant. The history of the improvement of an adsorption unit is shown in Table 1.

Table 1. Steps to improve the performance of an existing adsorption unit

	Initial conditions	1st retrofitting (enhancement of vacuum pump capacity)	2nd retrofitting (installation of membrane stage)	3rd retrofitting (enhancement of blower capacity)
Emissions	> 3 kg/h	3 kg/h	< 150 mg/m^3 goal, 20 mg/m^3 actual	< 150 mg/m^3 goal, 70 mg/m^3 actual
Electric power (installed)	64.5 kW	94.5 kW	131.5 kW	131 kW
Electric power (consumed)	57 kW	86 kW	118 kW	120 kW
VRU capacity	280 m^3/h	280 m^3/h	280 m^3/h	350 m^3/h
Specific power requirement	0.2 kW/m^3	0.31 kW/m^3	0.42 kW/m^3	0.34 kW/m^3

Figure 10. Gasoline vapor recovery unit. Interior view of a container unit (used with permission of Borsig Membrane Gas Processing)

The first step of plant improvement was the enhancement of vacuum pump size to make desorption more efficient and to improve the adsorption capacity. This measure led to emissions of 3 kg/h. To meet the stipulated goal of < 150 mg/m^3, a vacuum-assisted membrane stage was installed (Fig. 11). Higher hydrocarbons are separated efficiently by membranes, and the separated permeate stream is fed to the existing absorption unit for recovery. The stream with a residual content of organic vapor is finally treated by the adsorption unit. At the same feed volume flow the emission was reduced to 20 mg/m^3. This low emission allowed an increase of feed volume flow. Because of spare capacity of the blower at the outlet of the adsorption unit the feed volume flow could be increased from 280 to 350 m^3/h. The outlet concentration was still half of the allowed limit, and the specific power consumption of the plant could be reduced from 0.42 to 0.34 kW/m^3.

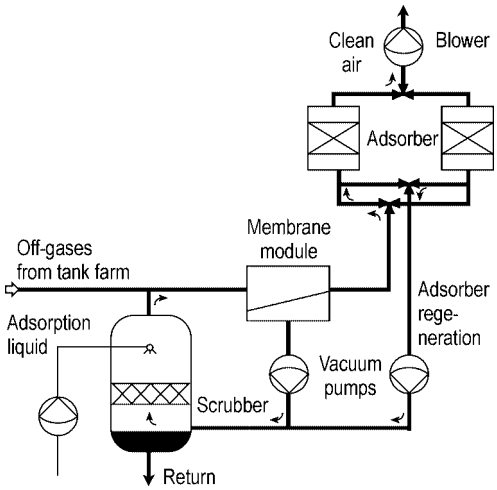

Figure 11. Retrofitting of an adsorption unit by installation of a membrane stage

4.3. Emission Reduction at Gas Stations

The end of the chain of gasoline distribution is the gas station. Emissions are generated by car refueling and from the stack of the storage tanks. The refueling emissions are reduced by returning the vapors by means of vacuum-assisted systems or balance systems. Vapors generated by car refueling at the filler neck of the car tank are collected by a device at the nozzle of the gasoline dispenser. Stack emissions are caused by a higher vapor return rate, evaporation of unsat-

urated air, temperature change, changes in ambient temperature, and refilling of the storage tanks. To avoid emission transfer from the car tank neck to the storage tank stack an air to liquid ratio of 1:1 has been stipulated. The capture of the refueling emissions is highly dependent on the design of the car tank neck. Investigations by TÜV Rheinland have shown that the capture efficiency is limited to an average of 75 % [15]. Meanwhile, neck designs with optimized and improved capture rates have been demonstrated. Tests have shown that an increase of the vapor return ratio from 1:1 to 1.5:1 leads to a higher vapor capture efficiency of 95 – 99 % [16]. Enhanced vapor return rates are only allowed if a vapor recovery unit with efficiency higher than 97 % is installed in order to separate and recover the gasoline vapors [14].

An essential requirement is leakage-proof installation of tanks, pipes, and dispensers. The storage tank stocks must be equipped with pressure/vacuum valves and the tank-filling pipes and vapor balancing pipes with check valves to avoid uncontrolled emissions. The system for the control of stock emissions (Fig. 12) consists of a membrane module (6), vacuum pump (7), pressure gauge (9), and a pneumatic control valve (10). The pressure gauge measures the pressure difference between tank head space and atmospheric pressure. If the pressure difference reaches a given set point, the vacuum pump (7) is started, the control valve opens by means of the applied vacuum at the permeate side, and the tank pressure is released by passing the head space vapors over the membranes arranged in the membrane modules. The gasoline vapors are separated from the vent stream and returned to the head space of the storage tank.

4.4. Off-Gas Treatment in the Chemical Industry

Dichloromethane (DCM) is a solvent which is commonly used in chemical production. DCM is considered to be a hazardous compound. The allowed effluent concentration is 20 mg/m^3. Figure 13 shows a flow scheme of a DCM separation plant for the treatment of highly concentrated off-gas streams. The operating conditions and relevant information on process data are listed in Table 2.

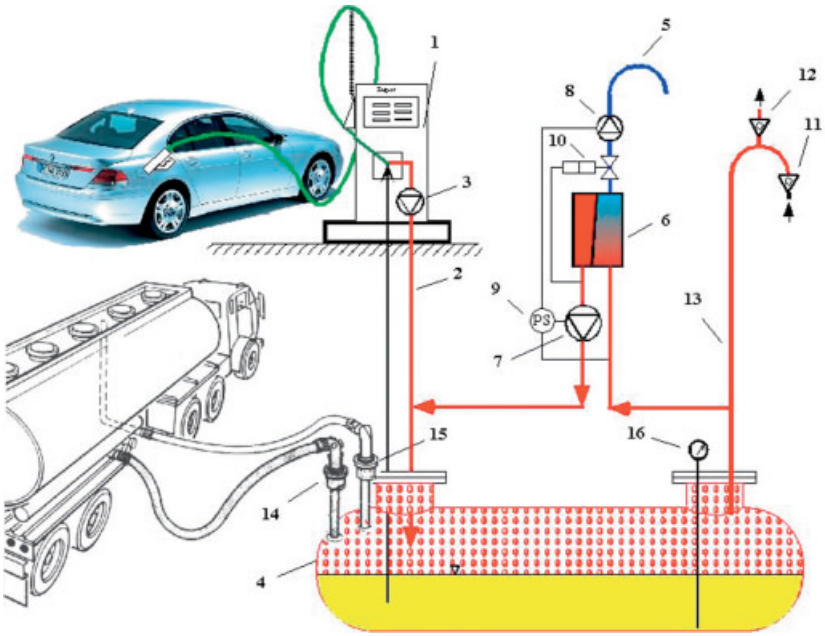

Figure 12. Vapor recovery at gas stations
1) Petrol pump with vapor return nozzle; 2) Vapor return nozzle; 3) Vapor return pump; 4) Underground storage tank; 5) Stack; 6) Membrane module; 7) Vacuum pump; 8) Retentate pump; 9) Pressure switch; 10)Pneumatic valve; 11) Air intake valve; 12) Air outlet valve; 13) Breather pipe; 14) Product fill connectors; 15) Vapor adaptor; 16) Electronic level guage

Table 2. Operating conditions for DCM recovery

Operating point	1	2	3	4	5	6	7	8	9
Temperature, °C	25	20	5	10	10	20	20	20	20
Pressure, bar	1	1	5	4.7	0.1	4.6	0.1	0.1	0.1
Feed volume flow, $m^3{}_{STP}/h$	75	115	100	72	28	40	72	12	40
DCM content, vol %	20	16.8	4.9	0.047	15.3	10.6	< 0.0001	0.3	10.6
DCM, $g/m^3{}_{STP}$	758	709	185	1.8	579	400	< 0.020	11	400

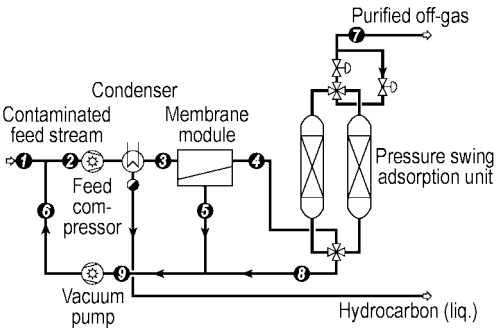

Figure 13. Flow scheme of a DCM recovery plant

Membrane separation can be advantageously applied to treat concentrated gas and vapor streams. The membrane is often used as prestage to optimize recovery by adsorption or condensation. Figure 14 shows a combination of adsorption and membrane technology to treat low-concentration off-gas streams. The off-gas from a production plant is contaminated with low concentrations of DCM and ethanol (EtOH) and is saturated with water vapor.

The off-gas is cooled to reduce the water content before a booster feeds the stream to the adsorption stage. Two parallel arranged adsorber units are installed to treat the off-gas. One unit is in the adsorbing phase, whilst the other unit is in the regeneration phase. The cycle time is approximately 10 min. Regeneration is facilitated by applying a vacuum of 100 mbar and purging the column with a fraction of cleaned air from the unit in adsorption operation. The DCM- and EtOH-enriched desorption stream is compressed to 6 bar and the organic vapors are liquefied at 15°C. Residual organic vapors are separated by the membrane stage after the condensation step and recycled to the liquid ring compressor for recovery. The gas leaving the membrane stage is fed to the adsorption stage in operation for final purification. Technical data for the process follow:

Off-gas intake	$450\,m^3{}_{STP}/h$
DCM concentration	max. $20\,g/m^3{}_{STP}$
EtOH	max. 10
Purge	$75\,m^3{}_{STP}/h$
Desorption pressure	< 100 mbar abs
Feed pressure	6 bar abs
Condensation temperature	15°C
Cycle time	ca. 10 min
Concentration purge to recovery	$< 1\,g_{VOC}/m^3{}_{STP}$
Residual DCM concentration	$< 20\,mg/m^3{}_{STP}$
Residual EtOH concentration	$< 150\,mg/m^3{}_{STP}$

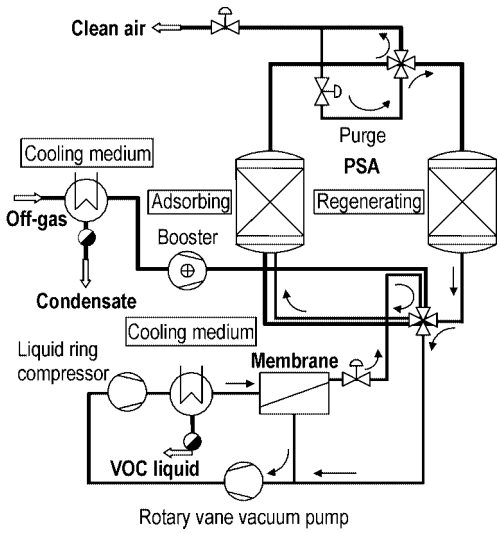

Figure 14. Hybrid system: First stage adsorption, second stage strip gas treatment by membrane separation (by permission of Satorius AG)

4.5. Recovery of Vinyl Chloride Monomer from Off-Gas of PVC Production

Vinyl chloride monomer (VCM) in off-gas from PVC production must be recovered because of its high commercial value. Residual concentrations from the recovery stage must be separated in view of its impact on the environment and

human health. The conventional vinyl chloride recovery process is depicted in Figure 15.

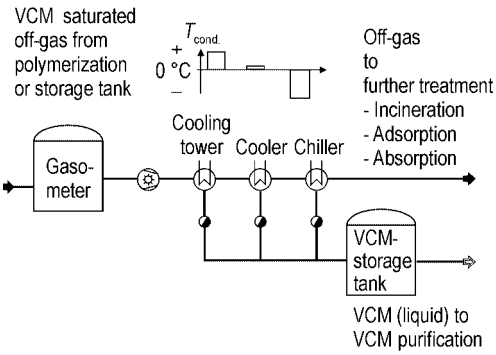

Figure 15. Conventional vinyl chloride recovery process

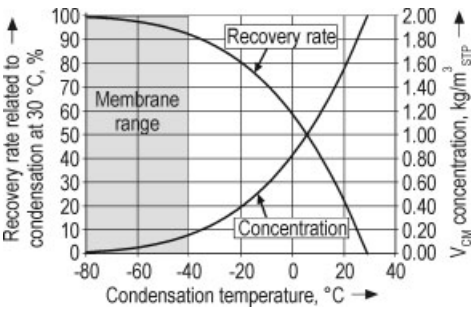

Figure 16. VCM recovery rate and residual concentration versus condensation temperature

The commonly used recovery process is compression of the off-gas to 5.5 bar and condensation in multistage condensers. The first condenser is operated with cooling tower water at ambient temperature, the second condensation is performed at –10 to –15°C and the third stage at approximately −40°C (92 % recovery). The residual VCM concentration leaving the recovery stage is 0.18 kg/m$^3_{\mathrm{STP}}$ (Fig. 16).

Besides the considerable residual concentrations another disadvantage is ice formation on the condenser surface when water vapor is present in the off-gas. Two parallel pressure condensation stages must be installed to perform a continuous recovery operation. The second and third condensation stages can be replaced by a membrane stage. Depending on the membrane stage layout separation efficiencies can be achieved which are comparable with those at condensation temperatures of −70°C. Figure 17 shows retrofitting of a multistage condensation unit.

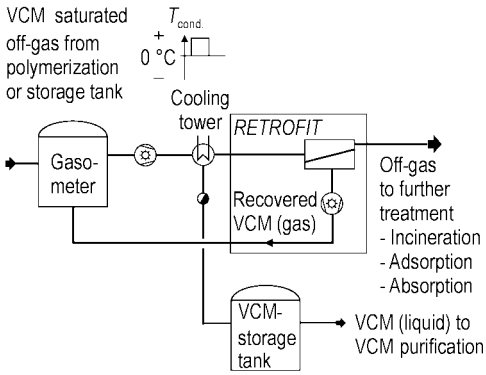

Figure 17. Retrofit of a multistage condensation unit. The separated VCM vapors are recycled to the gas holder and recovered in the first condenser operated with cooling water at ambient temperature

4.6. Batch Reactions and Vacuum Drying of Fine Chemicals

The drying process starts at atmospheric pressure. A speed-controlled dry vacuum pump evacuates the reaction vessel as fast as possible and maintains a constant low pressure at the permeate side of the membrane. A precondenser treats leakage air saturated with solvents to reduce the volume flow mainly in the beginning of the evacuation process. After compression by the dry vacuum pump solvents are partly condensed in the condenser before the stream is fed to the membrane stage and the remaining vapors are removed to a great extent. The solvent-enriched stream is recycled to the suction side of the vacuum pump, whereas the cleaned air stream is vented. Figure 18 shows a flow scheme of the process, and technical data are listed below:

Typical solvents	acetone, toluene, dichloromethane
Process pressure	atmospheric down to 1 mbar
Leakage air	up to 5 kg/h
Condensation temperature	−20 to +20°C (depending on solvent)
Solvents in vent stream:	0.1 kg/h (typical)
Power consumption	max. 10 kW at final pressure.

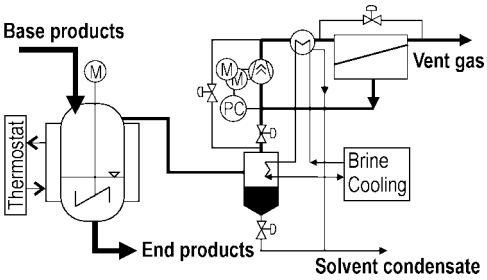

Figure 18. Flow scheme of vacuum drying of fine chemicals

4.7. Polyolefin Production Processes

One breakthrough of membrane-based separation of organic vapors was the introduction of membrane stages into continuous production processes. Areas of applications in polyolefin production are raw material purification, chemical reaction, and product purification, and finishing [17, 18]. Propylene, ethylene, and hexane are separated from process streams, and nitrogen can be purified and recycled for further use.

Propylene Recovery and Gas Treatment. The process objectives are the separation and recovery of propylene and the purification of nitrogen to be reused in the production plant. Propylene recovery is based on a process involving compression, refrigeration/condensation, and membrane separation.

The off-gas from the production plant is compressed by an oil-lubricated screw compressor. The discharge pressure is typically ca. 16 bar to provide sufficient driving force for the membrane separation process. The compressed process gas is released from the compressor discharge cooler at ca. 40°C and fed to a cryogenic condensation unit. A typical condensation temperature is −20°C. The residual noncondensed hydrocarbon compounds are separated by membrane separation into a hydrocarbon-depleted nitrogen stream at the high-pressure side (retentate) and a hydrocarbon-enriched stream at the permeate side. The membrane system is built as a multistage unit (Fig. 19). The first stage is typically operated with atmospheric pressure at the permeate side. To enhance the separation efficiency, vacuum can be applied at the permeate side of the subsequent stage. Finally, all permeate streams are combined to be recycled to the compressor suction.

The basic concept can be modified and the process can be tailored to specific process requirements. For instance, the number, arrangement, and operating parameters of the individual membrane stages can be adapted to optimize the overall process performance. It is also possible to employ a pressure swing adsorption unit for final polishing of the retentate stream if high nitrogen purities are required.

Polyethylene Powder Treatment. Membrane separation systems can be employed advantageously in the production of polyethylene for recovery of hydrocarbons (predominantly

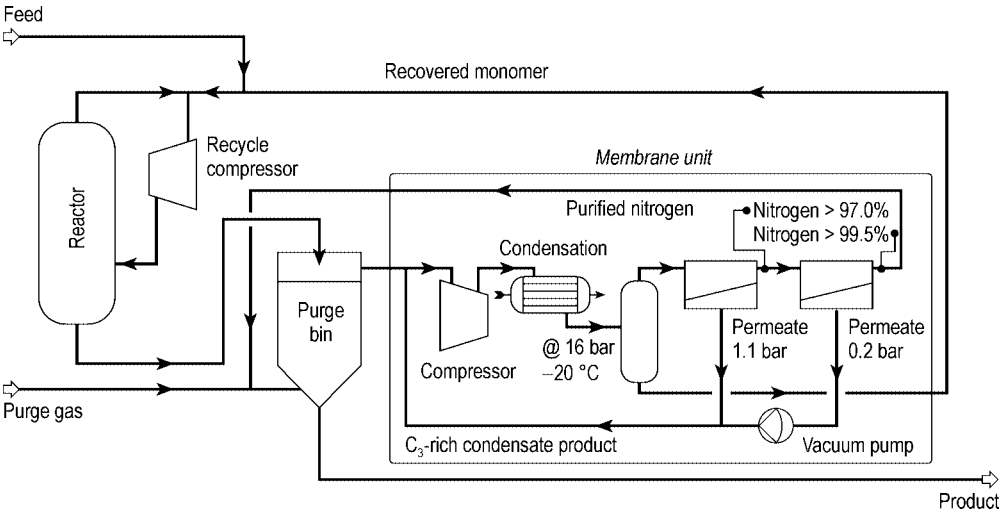

Figure 19. Membrane-based propylene recovery unit

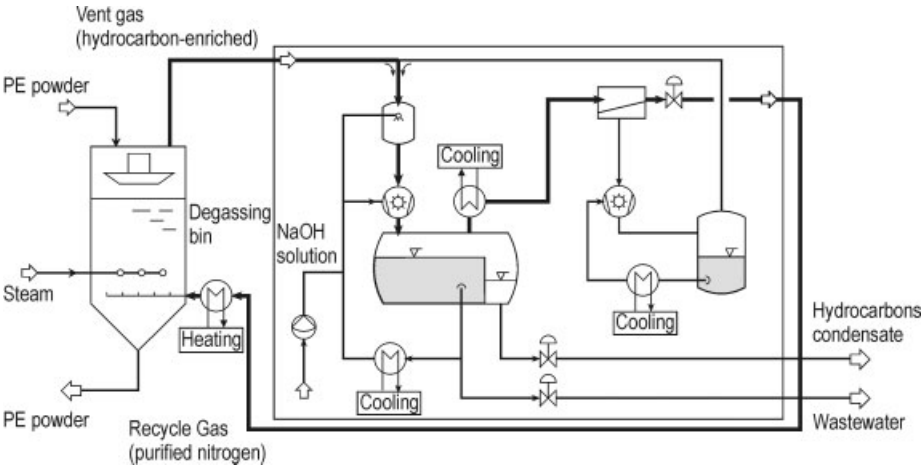

Figure 20. Flow scheme of PE powder treatment (used with permission of Basell Polyolefin GmbH)

hexane) and purification of recycle nitrogen. The membrane system is employed to treat the off-gas from the degassing bin. In this process step hydrocarbons are separated from the produced PE powder by purging with heated nitrogen. The requirements of the membrane process are:

- Purification of nitrogen for reuse as recycle gas to minimize off-gas treatment capacity.
- Recovery of *n*-hexane for reuse as feedstock in the polymerization process.
- Separation of water/acid from the recovered hydrocarbons for disposal in a suitable way.

Figure 20 shows a flowsheet of the process. Hydrocarbon-enriched vent gas from the degassing bin is fed to a quench cooler at atmospheric pressure. The cooler decreases the gas temperature from 100 to 40°C by spraying water into the suction line. The water supply is provided by the service liquid loop of the liquid-ring compression system. *n*-Hexane has negligible solubility in water. Hence, water is well suited as service liquid. The liquid-ring compressor installed downstream of the spray cooler increases the gas pressure to 7 $\mathrm{bar_{abs}}$. Subsequent to the compression stage the gas is cooled to 30°C and partly condensed. The condensate forms an aqueous and an organic (hydrocarbon) phase. The ring liquid tank also serves as phase separator from which the hydrocarbons are recycled to the overall process feedstock and water is fed to the liquid-ring compressor. The gas – vapor mixture exiting the condenser is fed via a droplet filter to the membrane modules. In the hydrocarbon-selective membranes hexane is separated from the feed gas by permeation through the membrane. The driving force of the permeation process is the partial pressure difference across the membrane. Hence, a liquid-ring vacuum pump operating at 200 mbar is installed on the permeate side to support transmembrane mass transfer. The permeate stream is recycled to the compressor inlet. The retentate stream exiting the high-pressure side of the membrane system is purified nitrogen. The stream is almost free of water, as water also permeates through the membrane and is concentrated on the permeate side. The purified nitrogen is reheated and fed to the degassing bin. Employing liquid-ring-type machinery allows for simple neutralization of the acidity of the incoming portion: aqueous NaOH solution is fed to the service liquid loop by a pH-controlled metering pump. A membrane stage for dehydration of process gas streams is shown in Figure 21. Technical data of the process follow:

Capacity	ca. 415 $\mathrm{m^3_{STP}}$/h at 1.05 bar and 100°C
Composition	ca. 500 kg/h nitrogen, 45 kg/h hydrocarbons (> 99 % *n*-hexane) 5 kg/h water
Condensation	ca. 30°C
Membrane	feed: 7 bar, retentate: ca. 6.5 bar, permeate: 0.2 bar
Power requirements	ca. 120 kW (compressor) +30 kW (vacuum pump)
Cooling water	ca. 35 $\mathrm{m^3}$/h at 25°C, $\Delta T = 5$ K
Outlet concentration	< 500 ppm hydrocarbons, < 300 ppm water
Recovery	> 99.99 % hydrocarbons

Figure 21. Membrane stage for dehydration of process gas streams (used with permission of Sterling-SIHI Anlagentechnik)

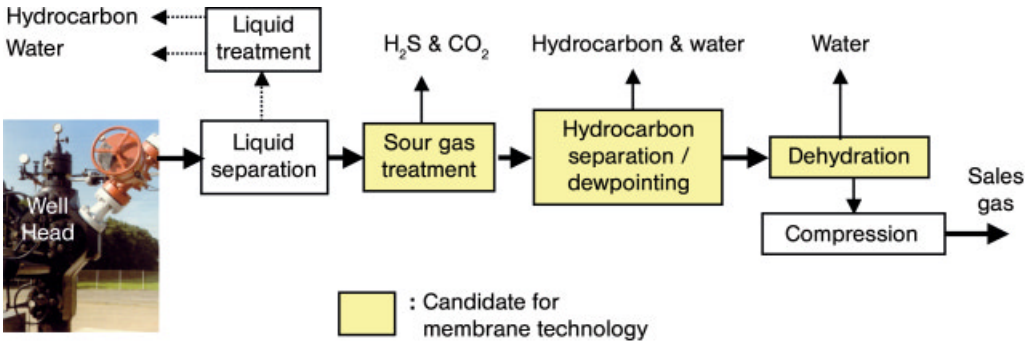

Figure 22. Processing steps of natural gas purification

4.8. Oil and Gas Production

Natural Gas Purification. Natural gas produced at the wellhead must be treated according to the required pipeline and quality specification. Figure 22 shows the gas processing steps and the potential for membrane applications. The separation of CO_2 from gas streams has already been realized in large-scale applications [19]. Hydrocarbon dew pointing and water-vapor dewpointing of natural gas are applications of interest and have been tested in pilot scale [20].

Based on the experience gained from gasoline vapor recovery, low-pressure applications of natural gas treatment have been realized. One example is the treatment of municipal gas. Gas supplied from a public utility company was used as chemical feed stock. Because of its high content of ethane and compounds the gas must be pretreated to enhance the methane content from 86.5 to >95 vol %. The applied technique was pressure swing adsorption. Because of the high costs for replacement of activated carbon and its waste disposal, alternative technologies have been considered. The pressure swing adsorption unit was replaced by a membrane stage, as shown in Figure 23.

A single-stage membrane was installed. The separation process is supported by applying vacuum at the permeate side. The permeate stream with higher hydrocarbon content was fed to a boiler for steam generation.

If no economic use for the permeate exists, a membrane process with permeate recycling can be designed. Process optimization results in a

Table 3. Feed, retentate, and permeate compositions

Feed volume flow, $m^3{}_{STP}/h$		35		126	
Stage cut		0.59		0.25	
Compounds	Feed, mol %	Retentate, mol %	Permeate, mol %	Retentate, mol %	Permeate, mol %
Methane	46.2	61.6	36.4	53.8	22.4
Ethane	9.73	5.3	12.36	9.51	11.1
Propane	11.1	2.72	16.09	8.56	20.3
Isobutane	2.63	0.48	3.91	1.82	5.36
n-Butane	5.78	0.64	9.14	3.39	13.0
Isopentane	3.04	0.15	5.25	1.32	7.43
n-Pentane	3.21	0.17	5.53	1.27	8.30
N_2	13.5	27.3	4.79	17.02	2.27
CO_2	1.82	1.14	2.25	1.83	1.80
H_2S	0.09	0.13	0.07	0.11	0.0
He	0.08	0.001	0.12	0.007	0.19
C_{6+}	2.62	0.15	4.02	1.09	7.78
Methane number	36	71		49	

lower requirement of feed gas at a given product volume but an increase of membrane area and the requirement of an additional vacuum pump and recycle compressor.

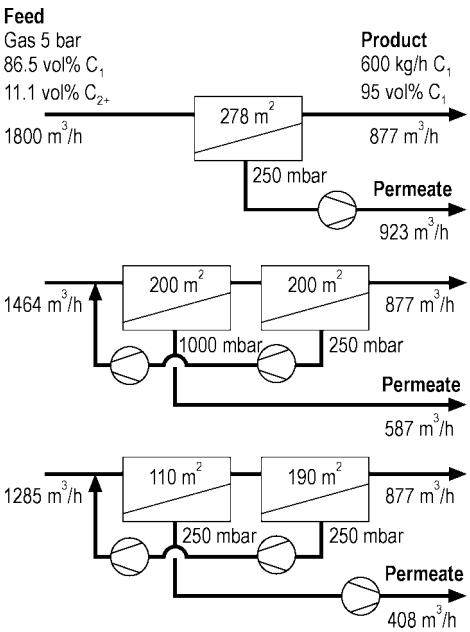

Figure 23. Treatment of municipal gas to meet chemical feedstock requirements

Methane Number Control and Treatment of Associated Gas from Oil Production. The performance of a gas engine operated with natural gas is dependent on the methane number of the feed gas. The methane number characterizes the antiknock behavior of the fuel gas and is analogous to the octane number of gasoline. Figure 24 shows a pilot plant consisting of a membrane module and a vacuum pump. This plant was installed in a bypass of a feed line to a gas engine.

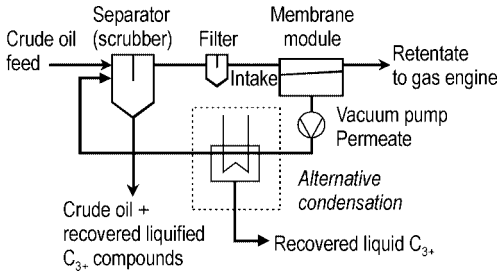

Figure 24. Simplified flow scheme of a membrane unit for methane number control

A methane number of 50 was sufficient for reliable operation of the gas engine. Table 3 shows the methane number enhancement at a feed pressure of 1.6 bar, a permeate pressure of 180 mbar, and a feed gas temperature of 27°C.

Membrane selectivity, installed membrane area, and available feed pressure are fixed values. Under the given conditions the increase in methane number is achieved by a reduction of feed volume flow, which is associated with an enhanced permeate flow because of the fixed membrane area. Stage cut and achieved methane number are in a direct correlation.

A comparison of alternative modes of operating the gas engine shows the following results.

Methane number	33 – 39
Calorific value of fuel gas	$H_s \approx 19\,\mathrm{kWh/m^3{}_{STP}}$
Fuel gas volume flow	$90\,\mathrm{m^3{}_{STP}/h}$
Max. power production	490 kW.

Methane number control by means of membranes:

Methane number	46 – 50
Calorific value of fuel gas	$H_s = 15\,\mathrm{kWh/m^3{}_{STP}}$
Fuel gas volume flow	$120\,\mathrm{m^3{}_{STP}/h}$
Max. power production	550 kW

To operate a gas engine with untreated associated gas the engine performance must be throttled to operate the engine in an acceptable antiknock range. The engine runs more efficiently and reliably if the methane number is increased. The engine performance and the power generation can be expanded by improved antiknock property of the fuel gas.

5. References

1. S. Verstraete et al.: "VOC Separation on Membranes: A Review" in *Characterization and Control of Odours and VOC in the Process Industries*, Studies in Environmental Science **61**, Elsevier, Amsterdam 1994, pp. 359 – 373.
2. S. P. Nunes, K. V. Peinemann: *Membrane Technology in the Chemical Industry, Part 1. Membrane Materials and Membrane Preparation*, Wiley-VCH, Weinheim 2001, pp. 6 – 38.
3. GKSS Forschungszentrum Geestacht GmbH, US 5 595 658, 1997.
4. K. Ohlrogge et al.: "Engineering Aspects of the Plant Design to Separate Volatile Hydrocarbons by Vapor Permeation", *Sep. Sci. Technol.* **28**, (1993) nos. 1 – 3, 227 – 240.
5. W. L. Robb: "Thin Silicone Membranes — Their Permeation Properties and Some Applications", *Ann. N.Y. Acad. Sci.* **146** (1986) 119.
6. J. A. Barrie, K. Munday: "Gas Transport in Heterogeneous Polymerblends". *J. Membr. Sci.* **13** (1983) 175.
7. R. M. Barrer, J. A. Barrie, N. K. Raman: "Solution and Diffusion in Silicone Rubber", *Polymer* **3** (1962) 595.
8. A. Alpers: Hochdruckpermeation mit selektiven Polymermembranen für die Separation gasförmiger Gemische. GKSS 98/E/21, ISSN 0344-96 29.
9. A. Alpers et al.: "Organic Vapor Separation: Process Design with Regards to High-Flux Membranes and the Dependence on Real Gas Behavior at High Pressure Applications", *Ind. Eng. Chem. Res.* **38** (1999) 3754 – 3760.
10. R. Rautenbach: "Gaspermeation", in: *Membranverfahren*, Springer Verlag, Berlin 1997, pp. 312 – 354.
11. GKSS Forschungszentrum Geestacht GmbH, DE 198 56 068, 2000 (J. Wind, K. Ohlrogge, B. Keil, A. Alpers).
12. *Technische Anleitung zur Reinhaltung der Luft (TA Luft)*, Weka Fachverlag, Kissing.
13. BImSchV–Verordnung zur Begrenzung der Emissionen flüchtiger organischer Verbindungen beim Umfüllen und Lagern von Ottokraftstoffen.
14. BImSchV – Verordnung zur Begrenzung von Kohlenwasserstoffemissionen bei der Betankung von Kraftfahrzeugen.
15. D. Hassel: Mess- und Überwachungsverfahren bei Gasrückführsystemen. *Praxis-Forum Umweltmanagement. 2. Gaspendel-Praxis-Seminar, 23/94. Technik* + Kommunikation Verlags GmbH, Berlin, pp. 183 – 203.
16. K. Ohlrogge, J. Wind: *Möglichkeiten der Anwendung von Membrantrennanlagen an Tankstellen. Praxis Forum Umweltmanagement. 2. Gaspendel-Praxis-Seminar*, 23/94. Technik + Kommunikation Verlags GmbH, Berlin, pp. 207 – 217.
17. R. W. Baker, M. Jacobs: "Improved Monomer Recovery from Polyolefin Resin Degassing", *Hydrocarbon Processing*, March 1996.
18. M. Jacobs, D. Gottschlich, F. Buchner: "Monomer Recovery in Polyolefin Plants Using Membranes — An Update" *1999 Petrochemical World Review*, D. Witt & Company, Houston, Texas, March 23 – 25, 1999.
19. S. A. Stern, W. J. Koros: "Separation of Gas Mixtures with Polymer Membranes. A Brief Overview", *Chim. Nouv.* **18**, (2000) 3201 – 3215.
20. S. Beauregard et al.: "A Novel Membrane Process for the Dehydration of Natural Gas", Proceedings 52nd Laurence Reich Gas Conditioning Conference 2002, Norman, Oklahoma, pp. 259 – 272.

Membranes: Gas-Separation Applications

DONALD J. STOOKEY, Elah Strategies, Chesterfield, Missouri, USA

1. **Introduction** 1
2. **Commercial Membrane Gas-Separation Applications** 1
2.1. **Hydrogen Separation** 1
2.2. **Helium Separation** 6
2.3. **Nitrogen Generation** 6
2.4. **Acid-Gas Separation** 8
2.5. **Gas Dehydration** 9
3. **Developing Membrane Applications** . 10
3.1. **Oxygen and Oxygen-Enriched Air** . . 10
3.2. **Nitrogen-Enriched Air (NEA)** 10
3.3. **Nitrogen Rejection from Natural Gas** 10
3.4. **Carbon Dioxide Removal from Liquid Hydrocarbons** 11
3.5. **Removal of Condensable and Volatile Organic Compounds (VOC)** 11
4. **References** 11

1. Introduction

The differences between the interactions of gas and vapor species with polymeric materials are used to accomplish gas separations and purifications with membrane devices. Figure 1 shows the relative flux rates for a variety of gases for a polyetherimide film.

The figure shows a five-decade range in the transport rates of different gases in this polymer and suggests that any number of the gas species might be separated from their mixtures due to the wide differences in their permeation rates for this polymer. These differences are unique for each polymer type. Some rubbery polymers have substantially higher fluxes than those shown in Figure 1, but these materials are typically less selective than glassy polymers. Membrane suppliers try to select polymers compatible with the application that will maximize the permeation fluxes and the selectivity for the gases involved.

Suppliers of membrane devices generally employ proprietary designs that direct the feed gas into intimate contact with a polymeric membrane surface across which the gases are separated. Besides the relative permeation rates of the species in the polymer, the feed composition, the driving force available for their separation, the membrane surface area, and the flow configuration also influence the degree of separation. The membrane device typically involves a pressure containment vessel and means for capturing permeate and retentate in ducts leaving the device. Examples of designs employed in gas separations are shown in Figures 2, 3 and 4. In Figures 2 and 3 the membrane is in the form of a flat sheet while Figure 4 depicts a hollow-fiber membrane device.

Table 1 summarizes applications and membrane suppliers. The key features of these applications are discussed below.

2. Commercial Membrane Gas-Separation Applications

2.1. Hydrogen Separation

Hydrogen is one of the more readily membrane separated gases. It has reasonably high selectivity in most glassy polymers relative to the other gas species (Fig. 1). It is also usually present in pressurized mixtures arising from the high hydrogen partial pressures employed to promote hydrogen's reactivity. Fortunately, high partial pressures are also essential for driving membrane separations. The cost of hydrogen manufacture is closely related to energy prices, since most hydrogen is produced by reforming of natural gas and hydrocarbons. There has been an ever-increasing demand for hydrogen owing largely to the higher severity in processing of heavier, sour crude oils while simultaneously reducing the sulfur content of fuel products to meet environmental regulations. Thus, the combination of technical feasibility plus environmental

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.j16_j02

Table 1. Membrane gas separations and membrane suppliers

Gases separated / Application	Air Products Prism Membranes	Air Liquide/MEDAL	Aquilo Gas Separation bv	Compact Membrane Systems	Cynara Membrane Separation Systems	Enerfex	Generon/Innovative Gas Systems	GKSS / SIHI Anlagentechnik	Kvaerner/Grace Membrane Systems	Membrane Technology & Research	Praxair/Innovative Membrane Systems	Sulzer Chemtech GmbH Membrane Systems	UBE Industries Ltd	UOP/Separex
H_2 from CO, N_2, CH_4														
CO and H_2/CO ratio adjusted synthesis gases	•	•											•	•
H_2 recovery from methanol process purge	•	•											•	
H2 recovery from ammonia process purge	•	•											•	
H_2 recovery from refinery process purge	•	•								•			•	•
High purity hydrogen for processes													•	
He from air and hydrocarbon gases														
Helium recovery and production from natural gas													•	
Purification of air contaminated helium	•												•	
CO_2 from CH_4, hydrocarbons														
Natural gas processing	•	•			•				•				•	•
Methane recovery from landfill gas and biogas		•							•				•	•
Carbon dioxide recovery for EOR flooding	•	•			•				•				•	•
Carbon dioxide removal from natural gas liquids					•									
CO_2 from N_2														
Carbon dioxide recovery from flue gases						•								
H_2S from hydrocarbons														
Sour natural gas processing and fuel conditioning	•	•			•					•				
O_2 from N_2														
Nitrogen generation from compressed air	•	•	•				•				•		•	
Nitrogen enriched air from compressed air				•										
Oxygen enriched air from compressed air				•										
N_2 from natural gas														
Gas conditioning for fuel cell						•								
H_2O from air and vapors														
Compressed air drying	•										•		•	
H_2O from hydrocarbon gases														
Natural Gas dehydration	•							•	•	•			•	
Alcohol and solvent dehydration												•	•	
Hydrocarbon vapors from gases														
Monomer and solvent vapor recovery								•		•		•		
Volatile organic compounds recovery from vent gases				•				•		•	•			

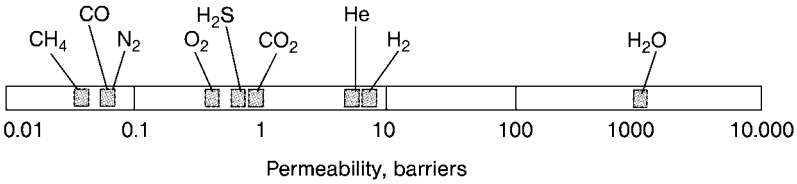

Figure 1. Permeability of gases through a polyetherimide film. 1 barrier = 10^{-10} $cm^3{}_{STP}$ $cm\,cm^{-2}\,s^{-1}cmHg^{-1}$

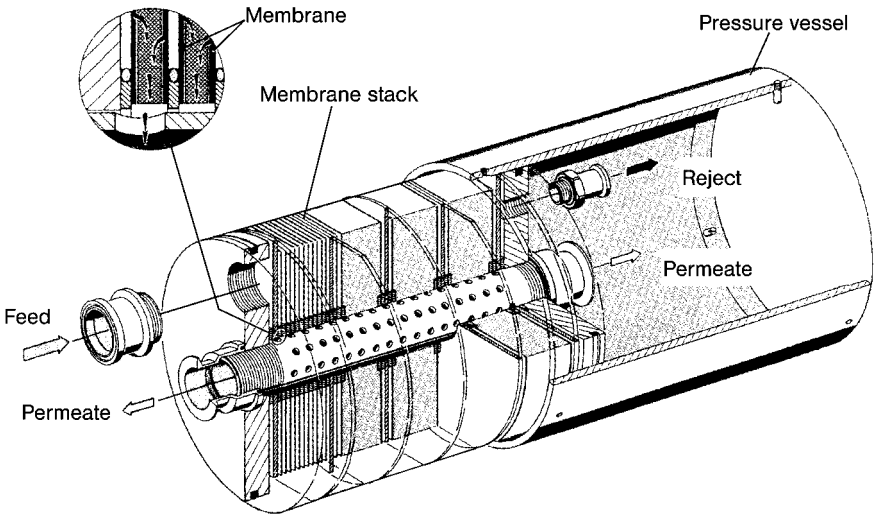

Figure 2. Flat-film membrane stack (used with permission of GKSS)

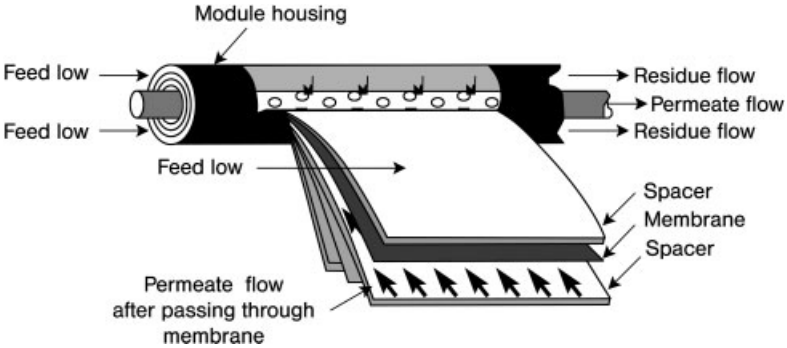

Figure 3. Spiral-wound module (used with permission of Membrane Technology & Research, Inc.)

and economic drivers has made hydrogen separation a major application of membranes.

Synthesis gases formed by steam reforming typically have hydrogen/carbon monoxide ratios on the order of 3/1. Adjustment of this ratio to 1/1 to satisfy the requirements of a carbonylation process [1] was the first commercial application of Monsanto's Prism Membrane technology. The application simply involves passing the 3/1 feed mixture across the membrane while withdrawing the hydrogen-rich permeate at reduced pressure. Control of the 1/1 hydrogen/carbon monoxide ratio in the carbonylation synthesis gas is accomplished under automatic process control by simply adjusting the membrane area, the feed rate, and the permeate pressure.

Hydrogenation of unsaturated hydrocarbons and hydrotreating to remove sulfur from fuels are two major consumers of hydrogen. Hydrogen supplied to reactors in these processes typically contains low concentrations of methane.

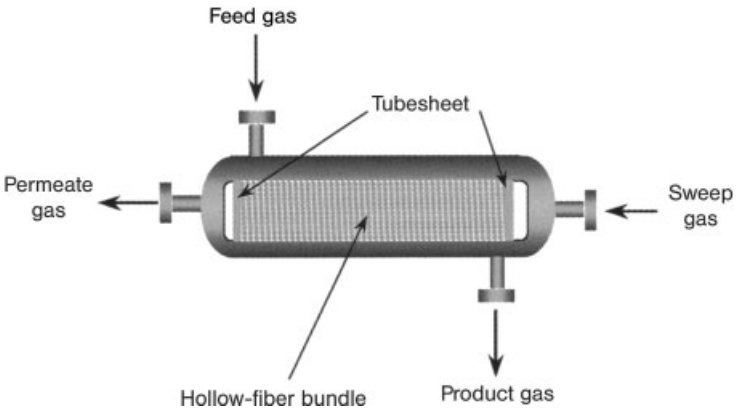

Figure 4. Axially oriented hollow-fiber module with bore and shell feeds (used with permission of Air Products and Chemicals, Inc.)

Methane and other inerts formed during hydrogenation and hydrotreating accumulate in the reactor and reduce the hydrogen partial pressure and hence the rate of the hydrogenation reaction. Thus, it is necessary to bleed off a quantity of the hydrogen reactant to purge the reactor of these inerts. Directing these pressurized purge gases to a membrane system enables a large fraction of hydrogen lost in the purge stream to be recovered as permeate and recycled to the reactor system. Inerts are thus rejected from the process at a higher concentration. The membrane system may also enable the hydrogenation to be optimized to operate at higher hydrogen partial pressures by increasing the reactor purge rate without a severe penalty on hydrogen yield. Indeed the scheme frequently enables improved throughput or quality of the hydrogenated product.

Ammonia synthesis reactions are also frequently limited by the accumulation of inerts in their reactor systems. Argon, arriving with the nitrogen reactant from air, and methane, arriving with the hydrogen, accumulate as inerts in the ammonia reactor loop. The purge from the ammonia loop is typically available in excess of 10 MPa pressure and contains more than 60 % hydrogen. Membrane systems processing this high-pressure purge typically recover over 90 % of the hydrogen. These systems typically employ two membrane units operated in series. The first recovers and recycles a large fraction of the hydrogen permeate to the final compression stage in the feed compressor of the ammonia plant, typically about 7 MPa. In the second membrane unit, the permeate is recovered at lower pressure and is recycled to the intermediate compressor stage, typically around 2.5 MPa. The rejected gas containing the concentrated inerts is typically burned as fuel in the primary reformer. The membrane systems enable ammonia plants to be optimized around higher productivity or energy savings [2 – 4]. The concentrated inerts can also be further processed by cryogenic or adsorption techniques for argon recovery [5, 6].

Methanol synthesis also deals with the accumulation of inerts in the reactor loop. In this case, the reformed gases supplied to the methanol reactor contain hydrogen, carbon monoxide, carbon dioxide, and small amounts of methane. The first three components react when the pressurized gases are circulated over the synthesis catalyst to form methanol and water. By cooling the reactor effluent gases, methanol and water are condensed and removed, leaving the methane to concentrate in the reactor circulation loop. By removing a small purge from the loop, the concentration of methane is controlled at the expense of the reactants also vented in this stream. These purged gases are available at reactor-loop pressure, typically 6 – 10 MPa, and thus have sufficient partial pressures of hydrogen and carbon dioxide to drive a membrane recovery unit. The purge gas is also saturated with methanol, a slow-permeating species that must be removed before the membrane system to avoid condensation when it is concentrated by removal of the fast-permeating species. This is accomplished in a water-scrubbing tower, the effluent of which can be combined with the crude methanol for

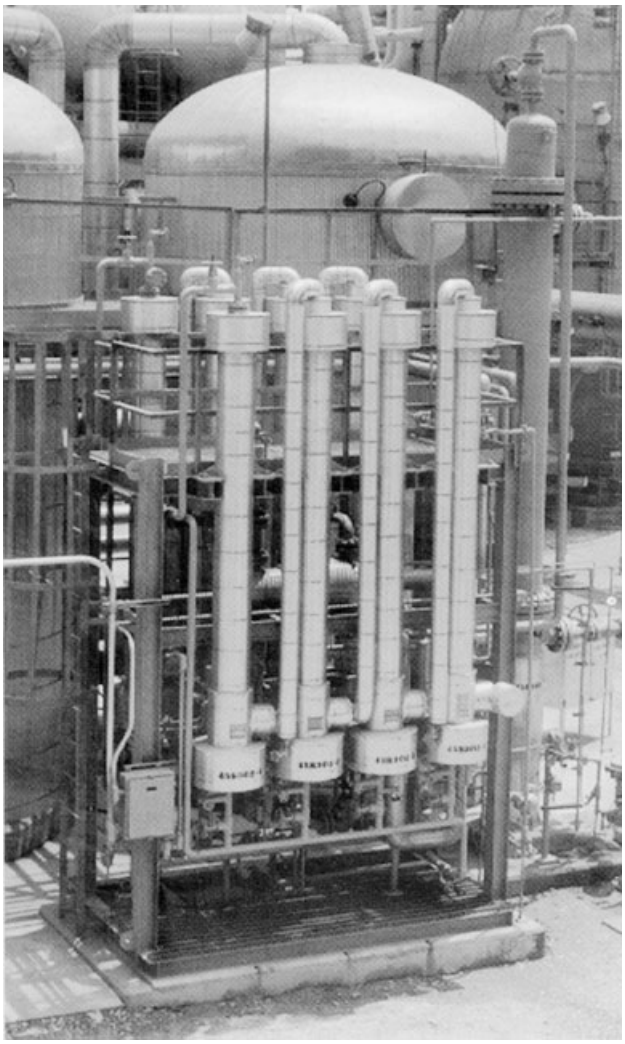

Figure 5. System for the recovery of methanol from hydrogen purge gas showing water-scrubbing tower (right) and membrane elements (center) (used with permission of Air Products and Chemicals, Inc.)

product distillation. Hydrogen and carbon dioxide are then recovered from the scrubbed gases in a membrane unit and recycled to the synthesis gas compressor. Figure 5 is a photograph of a purge-gas methanol recovery unit. Carbon monoxide is a slow-permeating species and therefore concentrates with the methane in the membrane reject stream that becomes fuel. Thus, as with ammonia purge recovery, the membrane unit provides efficient recovery of valuable reactants from the purge stream and provides the methanol producer with increased productivity and flexibility in plant optimization [7]. Starchem has proposed methanol production from a synthesis gas that is rich in nitrogen and contains only about 90 % of the stoichiometric amount of hydrogen needed to convert the carbon monoxide and carbon dioxide. They propose the use of a membrane unit to recover and recycle hydrogen to improve the reactant stoichiometry [8].

Hydrogen recovery systems have found similar utility in oil refining. In addition to recovery of valuable hydrogen [9], the hydrogen separation system often enables the refiner to operate at higher purge rates to increase the hydrogen partial pressure in the hydroprocessing unit. This can have a significant impact on catalyst life-

time and extend run times. It also can increase the unit's throughput and product quality. Hydrogen recovered from fuel streams can also be used to supplement hydrogen production, frequently delaying or deferring the need for added hydrogen production capacity.

Membrane processing of hydrogen-rich fuel gases can also improve the quality of the fuel gas, since the heat content of hydrogen is about one-third that of methane and higher hydrocarbons. Thus, by removing hydrogen, the heating value of the fuel gas can be adjusted to permit use of the fuel in conventional burners and equipment sized for the heating content of natural gas or even sold as pipeline fuel.

Hydrogen separation membranes are also frequently used in conjunction with other processing and separation technologies. Production of high-purity hydrogen by pressure swing adsorption (PSA) results in a waste stream containing over 40 % hydrogen. Frequently, additional hydrogen and fuel values can be realized by compression of the PSA purge and processing in a hydrogen membrane system [10, 11]. Removal of the hydrogen from hydrocarbon-rich fuel streams also enables recovery of condensable hydrocarbon liquids by processing the hydrogen-lean fuel gases in conventional expander plants.

2.2. Helium Separation

Like hydrogen, helium has high selectivity against most other components present in its mixtures for a wide variety of polymers. The primary source of helium is natural gas reservoirs, usually in concentrations below 1 %. Since low concentrations translate to low partial pressures, membranes find limited application in helium production, since multiple stages are required to reach even crude helium purities.

Membranes have found application in recovery of helium gas. Deep-diving gases employ helium mixtures that become contaminated. A combination of membranes and adsorbents are employed in purification of these gases for reuse. Small membrane systems are also employed in the purification of helium used in blimps. With time, the helium gas that inflates these lighter-than-air craft becomes contaminated with atmospheric gases. These contaminants are readily rejected by operating a small compressor and membrane unit to reject them and return purified helium to the craft.

2.3. Nitrogen Generation

Generation of nitrogen and nitrogen-rich atmospheres has become one of the largest uses for gas-separation membranes. The process simply involves passing filtered compressed air across the membrane. Since oxygen permeates faster than nitrogen, it is driven across the membrane by its partial pressure and concentrates in the permeate stream, leaving reject product enriched in nitrogen. The degree of oxygen removal is simply controlled by the residence time the nitrogen product has to contact the membrane. By reducing the product draw rate from the membrane, there is a correspondingly higher degree of oxygen removal. Thus, the oxygen concentration in the nitrogen-rich gas can be simply controlled by a process controller adjusting the product delivery rate to meet the specified oxygen concentration. Oxygen control levels vary with the application but typically fall within the 0.5 – 2.0 % range. These can typically be achieved in a single countercurrent membrane element. Concentrations below 0.1 % typically require two or more membrane elements in series [12]. Argon, carbon dioxide, and water vapor also transport readily across the membrane. Thus, the membrane-produced nitrogen is also very dry with measured dew points below −80°C.

Most nitrogen generators process compressed air at 0.7 – 2 MPa pressure and deliver the nitrogen-rich product within 0.1 – 0.2 MPa of the feed pressure. The degree of pretreatment of the air supplied to the membrane varies with the membrane employed and manufacturer. The performance of some membranes is sensitive to the moisture content of the air. In these cases, the compressed air is dried before processing. Other membranes are sensitive to organic vapors and oils, which are removed by carbon filters or absorption beds. Most generators are also supplied with filtration systems to remove condensate, mist, and any particulates that might plug the membrane flow channels or fiber bores. Since the permeability and selectivity of the membranes can be highly temper-

ature dependent, most applications require the temperature of the air feed to be controlled in order to control the quality of the product nitrogen or the membrane performance. The nitrogen product is typically delivered under pressure and flow regulation to a receiver from which it is dispensed to users on demand. The coproduced permeate is typically simply vented back to the atmosphere. The primary utility for the nitrogen is as an inerting atmosphere, typically for fire suppression. The critical oxygen concentration, the minimum level of oxygen required to sustain combustion, is typically between 9 and 12 % for many hydrocarbons and organic compounds. Thus, maintaining oxygen below 3 % provides ample safety margin for many inerting applications. These levels can readily be reached with membranes having oxygen/nitrogen selectivities greater than 4. Thus membrane-generated nitrogen finds ready application anywhere flammable materials are stored, processed, or handled. Storage in a low-oxygen atmosphere also preserves many fruits, vegetables, and produce. A low-oxygen atmosphere, typically in the 2 – 10 % range, depending upon the produce, is required. This is readily accomplished by purging the storage atmosphere with nitrogen or a nitrogen-rich gas as supplied by a membrane generator.

Membrane-based nitrogen generation equipment is supplied by a number of manufacturers (Table 1). Most are industrial gas suppliers who also use the equipment to supply smaller gas sales accounts. Membrane units are also used to supplement nitrogen supply at larger scale of gas accounts. Frequently, the membrane units are sized to handle a continuous, base-load requirement with truck-supplied liquid nitrogen providing backup and covering peak demand. Many of the membrane manufacturers also market membrane modules to original equipment manufacturers (OEMs). The OEMs incorporate the membrane modules and nitrogen generation capability into their own packaged equipment offers. Since compressed air is necessary for nitrogen generation, many OEMs are also compressor manufacturers and air-compressor distributors. Other OEMs are specialty equipment manufacturers whose equipment is dependent on a nitrogen supply. For example, a manufacturer of metal-processing furnaces often incorporates inert-atmosphere generation as part of his offer. Similarly, an instrument manufacturer may incorporate a small nitrogen generation system in his package to free his product from dependence on laboratory or cylinder nitrogen supply. Their features of portability, light weight, compactness, and robustness have aided in their incorporation into nitrogen-generation packages that literally can operate on land, at sea, or in the air. Some of the applications served by membrane nitrogen-generation equipment follow.

Shipboard Inert Gas. Nitrogen generated onboard ships is used to purge the vapor space in vessels containing liquid chemicals or flammable hydrocarbons on chemical tankers, the seals on cryospheres on LNG tankers, and various electronic systems on navy ships.

Controlled Atmosphere Storage. Membranes are used to maintain controlled atmospheres in warehouses for perishables, cargo containers, and the holds on banana reefer ships.

Tank and Process Equipment Inerting. Many offshore platforms, marine terminals, refinery and chemical tank farms are supplied with membrane-generated nitrogen.

Industrial Gas Sales. Many membrane suppliers are owned by or affiliated with industrial gas companies. They use on-site nitrogen generators to supply nitrogen to meet customer demand from base load to total-site nitrogen requirements, frequently supplemented with liquid nitrogen supply or backup.

Laboratories, Photo Processors, etc. Many facilities have replaced nitrogen cylinders with small wall-mounted cabinets that generate nitrogen for laboratory use from compressed air.

Metals Processing. Membranes generate nitrogen for inerting many sintered-metal process furnaces, soldering station blanketing, and the like.

Tire inflation. Flight line ground support carts generate and store high-pressure nitrogen for inflating aircraft tires and struts. Tire molding operations and truck and automobile service centers are supplying compressed nitrogen in place of compressed air for tire inflation to reduce tire oxidation and improve tire life.

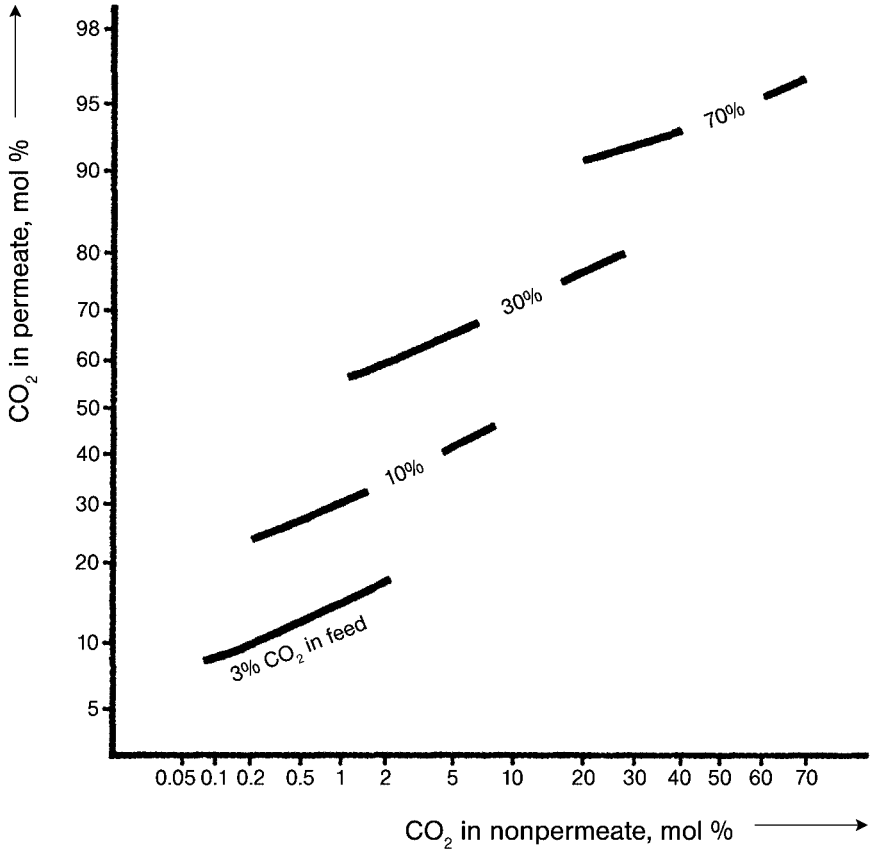

Figure 6. Typical membrane performance for acid gas removal from mixtures with methane or nitrogen (used with permission of Air Products and Chemicals, Inc.)

Oil and Gas Field Service. Membrane-produced nitrogen is replacing air in pneumatic drilling, used in oil well servicing, and for pipeline purging.

Flare and Seal Purging. Small nitrogen generators supply seal gas to a variety of seals on rotating machinery. They also supply purge to flare piping systems.

On-Board Inert Gas Generation System (OBIGGS). In the U.S. Air Force F-22, a small bleed of compressed air from the jet engine's turbine compressor is processed into nitrogen to purge the fuel tanks onboard the aircraft. OBIGGS is gaining increasing interest for military and commercial aircraft with the discovery of a fuel tank explosion as the probable cause of the TWA Flight 800 disaster.

2.4. Acid-Gas Separation

Figure 1 shows that the acid gases carbon dioxide and hydrogen sulfide have permeabilities in polyetherimide film that are about 20 times those of methane, nitrogen, and carbon monoxide. The latter are species that frequently accompany acid gases in mixtures of commercial interest. Commercial acid gas removal applications include production of pipeline-grade natural gas [13, 14], carbon dioxide recovery and recycle in enhanced oil recovery (EOR) [15 – 17], methane recovery from landfill and biogas [18, 19], and carbon dioxide recovery from flue gases [20, 21].

The performance of a typical acid gas removal membrane is shown in Figure 6. This graph provides operating lines for several feed compositions and predicts combinations of permeate and nonpermeate product composition

that can be produced when the membrane feed-to-permeate pressure ratio is greater than 20. Inspection of this diagram reveals the dilemma associated with membrane applications for acid gas separations, namely, that it is difficult to produce high purity products in either the permeate or the nonpermeate at high recoveries with a single membrane stage. Thus, it is necessary to have an alternative application for the impure coproduct, or resort to a more complicated process involving compression of the permeate to feed successive membrane stages to achieve high degrees of separation [22]. Study of the diagram reveals that processing a typical raw natural gas to a pipeline-grade natural gas containing less than 2 % carbon dioxide while rejecting carbon dioxide at greater than 90 % purity would require at least three membrane stages with permeate compression between each stage. The economy for such makes it practical in only a limited number of applications.

Despite this limitation, membranes are finding increasing application in acid gas removal. The successful applications depend upon identifying utility for the low-purity coproducts in another service or application. For example, processing a feed with 20 % CO_2 to pipeline-grade natural gas produces a medium-heating-quality fuel that is used for power generation. In another case, a membrane unit produces a high-CO_2 permeate gas suitable for injection into an oil reservoir for enhanced oil recovery, while the coproduct is processed further in a conventional amine sweetening process.

2.5. Gas Dehydration

The capability of membranes to dry gases was recognized in the early hydrogen, acid gas, and nitrogen membrane applications. Indeed some nitrogen generation systems were marketed for their ability to produce a very dry product suited for purging of cryogenic and LNG equipment. It was recognized early that a membrane's productivity of dry gas was dependent upon the creation of sufficient permeate volume to maintain the very low partial pressure driving force associated with a product containing only parts per million of water. Monsanto introduced the Cactus membrane air dryer in 1987, which relied on a different, higher flux membrane on the dry-product end of the separator [23]. This allowed a portion of the dried product to be used as a countercurrent sweep gas to create a good partial pressure driving force for the permeating water along the dryer length. Other dryers have since been introduced by a number of suppliers that rely on a four-ported membrane design as shown in Figure 4 to provide a controlled amount of sweep gas.

Most membrane dryers such as the Cactus Membrane Air Dryer line shown in Figure 7 are marketed to OEMs and compressor and instrument distributors. These devices process filtered, compressed air to deliver dry, compressed air having dew points as low as −40°C at 1 MPa pressure.

Figure 7. Cactus membrane compressed air dryers (used with permission of Air Products and Chemicals, Inc.)

Using membranes for drying natural gas was suggested as early as 1984 [24], but it was 1996 before the concept became reality when Air Product's Permea organization introduced a swept dryer design that was capable of drying gases to pipeline specifications at absolute pressures as high as 7 MPa [25]. Since then several

demonstration units with capacities as high as 200 KNm^3/d have been installed.

The high-pressure dryer has also found application in high-pressure compressed air systems onboard U.S. Navy vessels [26]. These dryers are typically operated between compressor stages at absolute pressures of 30 – 70 MPa and the dry product returned to the next compression stage for final boost to the 20 – 33 MPa receiver pressure.

3. Developing Membrane Applications

3.1. Oxygen and Oxygen-Enriched Air

All nitrogen generators necessarily produce an oxygen-enriched air (OEA) coproduct, but it is generally vented to the atmosphere. This is probably because the OEA is at low pressure, varies in composition due to process controls on the nitrogen generator, and volumes are small in comparison to meaningful OEA application requirements. One commercial application generates OEA for use in intermittent catalyst regeneration. This application is unique though as there is also an on-site application for nitrogen generation. Membranes have also been employed to produce OEA for ozone generators. Most attempts have involved the OEA coproduct from compressed-air-driven nitrogen generators. Since ozone generation is proportional to oxygen concentration and is favored by atmospheric-pressure operation, feeding the OEA to an ozone generator would appear to be a logical application. However, the economics are highly dependent on recovering value from the nitrogen coproduct stream. The application involves feeding the membrane with dry compressed air, since any moisture present in the feed copermeates with the OEA and interferes with ozone generation.

Only a small amount of the oxygen supplied is converted in an ozone generator. In a typical application, the ozone-rich stream is processed in an absorber or contactor tower to decompose organics in waste streams. Vent gases leaving such equipment are rich in oxygen, and thus there has been interest in their recovery and recycle to the ozone generator. Air separation membranes have been operated in such a recycle arrangement to enrich the OEA to as high as 60 – 70 % oxygen. At these levels, the ozone generator becomes very effective and more than doubles in productivity over an air feed. Such is the good news. Unfortunately, any residual ozone in the vent gas recycle stream also attacks the polymer membranes. With time, even ppm levels of ozone have a cumulative, irreversible, degrading effect on the membranes and shorten their useful life. Design of guard beds, ozone-destroying catalysts, and the like to protect the membranes have been successful during normal operation, but control systems designed for low ozone levels were not sufficiently responsive in detecting and responding to overload and upset conditions before damage to the membranes in the demonstration unit.

OEA production from air is favored by vacuum-driven membranes operating at low oxygen recovery. In this case, the OEA purity is simply a function of the membrane's selectivity and the vacuum level at which the OEA permeate is withdrawn. Unfortunately, the high selectivity membranes employed in nitrogen generators have low flux and would require huge membrane areas at subatmospheric pressure. Presently, high-flux materials are being used, but OEA purities are limited to less than 30 % oxygen. There is much research activity in pursuit of higher flux, more selective membranes that will bear watching for this application.

3.2. Nitrogen-Enriched Air (NEA)

Compact Membrane Systems is actively pursuing NEA as a replacement for exhaust gas recycle (EGR) as an NO_x abatement measure on diesel engines. The CMS membrane with its high flux and low selectivity is capable of enriching the turbocharged engine air to 81 – 84 % nitrogen. Such levels have been demonstrated as being effective in substantially reducing NO_x in the diesel engine exhaust [27].

3.3. Nitrogen Rejection from Natural Gas

A number of membrane developers and suppliers are pursuing nitrogen rejection from natural

gas with high-flux, low-selectivity membranes. Modest nitrogen depletion (30 – 50 % nitrogen removal) in the pressurized natural gas has been demonstrated at a beta site with plasma-treated polydimethylsiloxane membranes supplied by Neomecs [28]. Permeate from the nitrogen depletion membrane unit is nitrogen-enriched. As in the cases of acid gas removal discussed above, this application also depends upon the disposition of this low-energy gas.

3.4. Carbon Dioxide Removal from Liquid Hydrocarbons

Cynara [29] has recently piloted a membrane process for removing carbon dioxide dissolved in liquid ethane. Minor amounts of carbon dioxide in the natural gas feed to a turbo-expander plant can accumulate in the natural gas liquid (NGL) products. The Cynara membrane pilot processed the reflux to a de-ethanizer column. Driving force for carbon dioxide removal was provided by routing fuel gas for a process-fired heater to the sweep-gas side of the membrane. Adjustment of the sweep flow rate enabled 40 – 50 % of the carbon dioxide to be removed, thereby maintaining the contaminant concentration below the 6 % specified in the liquid ethane product.

3.5. Removal of Condensable and Volatile Organic Compounds (VOC)

A number of membrane suppliers have introduced membrane products for recovery and removal of hydrocarbons, VOCs, and condensable organic compounds from process and vent gases. Silicone membranes can separate many condensable hydrocarbon and organic vapors from noncondensable gases such as methane, nitrogen, and air. This is enabling them to be piloted and employed in a growing number of small applications. GKSS's membrane is being used to recover gasoline and fuel vapors from storage tank vent gases. Membrane Technology and Research (MTR) has demonstrated the utility of a silicone membrane in NGL recovery and dewpoint control in natural gas processing and in fuel gas conditioning for gas turbines and engines.

Compact Membrane Systems is developing a VOC recovery process based on their perfluoropolymer membrane. Unlike silicone membranes, which transport the hydrocarbon liquids, CMS's copolymer of perfluorodimethyldioxole (PDD) and tetrafluoroethylene (TFE) allows the noncondensable gases to pass through the membrane and rejects and concentrates the condensable hydrocarbons.

4. References

1. US 4255591, 1981 (E. C. Makin, J. L. Price, Y. W. Wei).
2. Monsanto Co., US 4172885, 1979 (E. Perry).
3. Monsanto Co., US 4180553, 1979 (H. R. Null, E. Perry).
4. Monsanto Co., US 4180552, 1979 (T. E. Graham, D. L. Maclean).
5. BOC Group, Inc., US 4750925, 1988 (D. L. Maclean, R. Krishnamurthy, S. L. Lerner).
6. BOC Group, Inc., US 4752311, 1988 (D. L. Maclean, R. Krishnamurthy, S. L. Lerner).
7. Monsanto Co., US 4181675, 1980 (E. C. Makin, K. K. Okamoto).
8. Starchem, Inc., US 5472986, 1995 (C. P. van Dijk).
9. Monsanto Co., US 4367135, 1983 (L. G. Posey, Jr.).
10. Monsanto Co., US 4229188, 1980 (G. M. Intille).
11. Monsanto Co., US 4238204, 1980 (E. Perry).
12. Permea, Inc., US 4894068, 1990 (A. W. Rice).
13. T. E. Cooley, W. L. Dethloff, "Field Tests Show Membrane Processing Attractive," *Chem. Eng. Progr.*, Oct. 1985, 45 – 50.
14. US 4130403, 1978 (T. E. Cooley, A. B. Coady).
15. C. S. Goddin, "Pick Treatment for High CO_2 Removal" *Hydrocarbon Process.* May 1982, 125 – 130.
16. W. J. Schell, C. D. Houston, "Process Gas with Selective Membranes" *Hydrocarbon Process.* Sept. 1982, 249 – 252.
17. W. A. Bollinger, D. L. Maclean, R. S. Narayan, "Separation Systems for Oil Refining and Production," *Chem. Eng. Progr.*, Oct. 1982, 27 – 32.
18. "Recovering Methane At A Small Landfill," *Waste Age*, Nov. 1983, 28 – 30.
19. D. J. Stookey, K. Boustany, R. L. Kilgour, "Recovery of Methane from Biogas with Monsanto's Prism Separators" BioEnergy 84 World Conference and Exhibition, Gothenburg, Sweden, June 18 – 21, 1984.

20. US 5233837, 1993 (R. A. Callahan).
21. US 6085549, 2000 (P. A. Daus, C. R. Pauley, J. W. Koenst, F. Coan).
22. D. J. Stookey, T. E. Graham, W. M. Pope, "Natural Gas Processing with Prism Separators," *Environm. Progr.* Aug. 1984, 212 – 214.
23. US 4687578, 1987 (D. J. Stookey).
24. F. Fournie, J. P. Agostini, "Permeation: A New Competitive Process for Offshore Gas Dehydration" OTC '84 Proceedings, 1984 Offshore Technology Conference, Houston, TX, May 7 – 9, 1984.
25. D. J. Stookey, K. Jones, D. G. Kalthod, T. Johannessen, "Membrane Dehydrators — A New Alternative for Drying High Pressure Gases" Presentation at The 1996 Membrane Technology/Planning Conference sponsored by Business Communications Co., Inc., Oct. 29, 1996, Newton, MA.
26. T. Theis, S. Titus, "The Development of Permeable Membrane Air Dehydrators for the U.S. Navy" *Naval Eng. J.*, May 1996, 243 – 265.
27. Compact Membrane Systems website, www.compactmembrane.com.
28. R. A. Callahan, Enerfex Inc., personal communications to author, Oct. 2000. See www.enerfex.com
29. G. H. Gall, E. S. Sanders, "Membrane Technology Removes CO_2 from Liquid Ethane," *Oil Gas J*, July 22, 2002.

My Profile
WILEY InterScience
HOME
ABOUT US
CONTACT US
HELP
Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Marine Fuels

Standard Article

Klaus Reders[1] and Torsten Mundt[2]
[1]Shell Global Solutions GmbH (retired), Hamburg, Federal Republic of Germany
[2]Shell Global Solutions GmbH, Hamburg, Federal Republic of Germany

Copyright © 2005 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.j16_j03
Article Online Posting Date: July 15, 2005

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: HTML

Abstract

The article contains sections titled:

1. History
2. Specifications
3. Composition
4. Properties
5. Outlook

[Top of Page]

1. History

Marine fuels for diesel engines cover a wide range of different grades, ranging from gas oil fractions often referred to as “distillates” —and including a grade with properties similar to automotive diesel fuel— to heavy residual fuel oils containing refinery residues remaining after crude oil distillation. While the distillate fuels are mainly used in the smaller high-speed engines (speed higher than 1000 rpm), the residual fuel oils are used in the large-bore, low-speed cross-head engines (below 350 rpm) and also in medium-speed engines.

For a long period marine diesel engines were solely fueled with the distillate type of fuels. Only after World War II the potential use of heavy residual fuels was investigated. These fuels with densities up to 1000 kg/m^3 and even above and viscosities up to 700 mm^2/s at 50 °C were originally used only as boiler fuels for steam generation. The low cost of these residual fuels was an incentive to investigate their use also in the large marine diesel engines. However, initially the cost saving obtained was more than offset by the increased difficulty in fuel handling and treatment (heating and separation of unsuitable material) and even more significantly by the higher rates of engine wear and additional maintenance cost. This situation changed dramatically after the so called “oil crises” in the 1970s. After that time the fuel cost increased and also the gap in the differential cost between the higher quality distillate fuels and the residual fuels widened, followed by a general move to the cheaper heavier fuels especially in the low- and medium-speed marine engines.

The higher fuel cost required also an improvement in the overall efficiency of the marine propulsion systems. As the diesel engine offered more scope (by approaching 50% total efficiency) than the steam turbine, an increasing number of new ships were equipped with diesel engines fitted with the treatment systems (mainly heating systems and centrifuges) required for the use of heavy residual fuel grades.

[Top of Page]

2. Specifications

While initially only fuel specifications by the suppliers existed, the need of international specifications was recognized with the increased use of heavy fuel oils in diesel engines. For that reason the British Standard BS MA 100 was introduced in 1982, followed by the ISO Standard 8217 in 1987. Furthermore there is an ASTM Standard D 975 and a diesel engine manufacturers and users specification issued by the International Council on Combustion Engines (CIMAC) and published first in 1986.

The ISO Standard 8217 specifies four distillate fuels (see Table 1), the three lighter ones DMX, DMA and DMB, to be used without specific fuel treatment on board, and 15 residual fuels, which require heating to reduce their viscosity and also generally centrifuging to separate water and foreign matter before their use in the engine (Table 2).

Table 1. Requirements for marine distillate fuels

Characteristic	Limit	Category ISO-F-				Test method reference
		DMX	DMA	DMB	DMC	
Appearance		Visual		–	–	See 6.2[5]
Density at 15°C, kg/m^3	max	[1]	890.0	900.0	920.0	ISO 3675 or ISO 12185 (see also 6.3[5])
Viscosity at 40°C, mm^2/s[2]	min.	1.40	1.50	—	—	ISO 3104
	max.	5.50	6.00	11.00	14.0	ISO 3104
Flash Point, °C	min.	43	60	60	60	ISO 2719 (see also 6.4[5])
Pour point (Upper), °C [3]						
winter quality	max.	—	–6	0	0	ISO 3016
summer quality	max.	—	0	6	6	ISO 3016
Cloud point, °C	max.	–16[4]	—	—	—	ISO 3015 (see also 6.5[5])
Sulfur, % (m/m)	max.	1.0	1.5	2.0	2.0	ISO 8754 (see also 6.6[5])
Cetane number	min.	45	40	35	—	ISO 5165 (see also 6.7[5])
Carbon residue [micro method. 10% (V/V) distillation bottoms], % (m/m)	max.	0.30	0.30	—	—	ISO 10370
Carbon residue (micro method), % (m/m)	max.	—	—	0.30	2.50	ISO 10370
Ash, % (m/m)	max.	0.01	0.01	0.01	0.05	ISO 6245
Sediment, % (m/m)	max.	—	—	0.07	—	ISO 3735
Total existent sediment, % (m/m)	max.	—	—	—	0.10	ISO 10307-1
Water, % (V/V)	max.	—	—	0.3	0.3	ISO 3733
Vanadium, mg/kg	max.	—	—	—	100	ISO 14597
Aluminum plus silicon, mg/kg	max.	—	—	—	25	ISO 10478 (see also 6.8[5])

[1]In some geographical areas. there may be a maximum limit.
[2]1 mm^2/s = 1 cSt.
[3]Purchasers should ensure that this pour point is suitable for the equipment on board. especially if the vessel operates in both the northern and southern hemispheres.
[4]This fuel is suitable for use without heating at ambient temperatures down to –15°C.
[5]Referring to ISO Standard 8217:1996(E).

Table 2. Requirements for marine residual fuels

Category ISO-F-

Characteristic	Limit	RMA 10	RMB 10	RMC 10	RMD 15	RME 25	RMF 25	RMG 35	RMH 35	RMK 35	RMH 45	RMK 45	RML 45	RMH 55	RMK 55	RML 55	Test method reference
Density at 15°C, kg/m^3	max.	975.0	981.0		985.0	991.0		991.0		1010.0	991.0	1010.0	—	991.0	1010.0	—	ISO 3675 or ISO 12185 (see also 6.3[3])
Kinematic viscosity at 100°C, mm^2/s [1]	max.	10.0			15.0	25.0		35.0			45.0			55.0			ISO 3104
Flash point, °C	min.	60			60	60		60			60			60			ISO 2719 (see also 6.4[3])
Pour point (upper), °C [2]																	
winter quality	max.	0	24		30	30		30			30			30			ISO 3016
summer quality	max.	6	24		30	30		30			30			30			ISO 3016
Carbon residue, % (m/m)	max.	10		14	14	15	20	18	22		22		—	22		—	ISO 10370
Ash, % (m/m)	max.	0.10			0.10	0.10	0.15	0.15	0.20		0.20			0.20			ISO 6245
Water, % (V/V)	max.	0.5			0.8	1.0		1.0			1.0			1.0			ISO 3733
Sulfur, % (m/m)	max.	3.5			4.0	5.0		5.0			5.0			5.0			ISO 8754 (see also 6.6[3])
Vanadium, mg/kg	max.	150		300	350	200	500	300	600		600			600			ISO 14597
Aluminum plus silicon, mg/kg	max.	80			80	80		80			80			80			ISO 10478 (see also 6.8[3])
Total sediment. potential, % (m/m)	max.	0.10			0.10	0.10		0.10			0.10			0.10			ISO 10307-2 (see also 6.9[3])

[1] Annex C gives viscosity/temperature table. for information purpose only. 1 mm^2/s = 1 cSt.
[2] Purchasers should ensure that this pour point is suitable for the equipment on board. especially if the vessel operates in both the northern and southern hemispheres.
[3] Referring to ISO Standard 8217:1996(E).

[Top of Page]

3. Composition

The hydrocarbon composition of marine diesel fuels depends mainly on that of the crude oil used in the refinery (e.g., distillates and distillation residues have compositions controlled by the crude oil) but also on the secondary refinery process units, e.g., aromatics content when the highly aromatic catalytic cracker cycle oil is used as a component in the residual fuels. Whereas the distillate fuels consist mainly of alkanes and cycloalkanes (i.e., paraffins and naphthenes), alkenes (olefinic hydrocarbons), and aromatics, the residual fuels contain also asphaltenes. These are very large hydrocarbon molecules with a very high carbon/hydrogen ratio and have complex, mainly aromatic structures and molecular masses between about 5000 and 40 000. In a stable residual fuel, that meets the standards described in Chapter Specifications, the asphaltenes are colloidally dispersed. Depending on the crude oil, the oil molecules contain also oxygen, sulfur, nitrogen, metals such as vanadium, chromium, nickel, aluminum and other nonhydrocarbon atoms in low concentrations.

When a paraffinic crude oil is used to produce a residual fuel oil often an aromatic diluent (such as cat cracker cycle oil) must be added in the refinery to prevent agglomeration of the asphaltene molecules and avoid sludge formation.

Long-chain alkanes in residual fuels may be present to some extent in form of wax at ambient temperature. Care must be taken to avoid precipitation and formation of waxy sludges during storage.

The residual fuel oils may also contain impurities such as water, sand, rust from tanks and pipelines, and catalyst fines from refinery processes, which must be separated prior to the use in the engine to avoid unacceptable wear in fuel pumps, injectors, and cylinder liners. The chemically bound sulfur present in marine diesel fuels (in the heavier grades up to 5%) require the use of specific highly alkaline cylinder lubricants to avoid corrosive wear.

[Top of Page]

4. Properties

Distillate Fuels. The specification of the lightest distillate fuel DMX is basically similar to that of automotive diesel fuel, however, DMX has no density and no distillation limits. These parameters are indirectly controlled by viscosity and cloud point. The minimum ignition quality (cetane number) of DMX is lower than that of automotive diesel fuel and the maximum sulfur content is higher. This distillate fuel is also referred to as an emergency fuel, because it can be used without any treatment on board down to low temperatures. The remaining heavier distillate fuels DMA (also referred to as gas oil), DMB, and DMC, have higher maximum density, viscosity, and sulfur levels, whereas low-temperature fluidity and cetane number are reduced. The more viscous fuels are geared to the requirement of medium-speed engines with longer piston strokes, offering more time for the air-fuel charge to ignite and burn during the combustion cycle and hence the ignition performance (characterizing ignition delay after injection and ease of combustion) of the heavier fuels is less important. The heaviest distillate fuel DMC may contain traces of residual fuel oils and therefore specification criteria used in that fuel category to control impurities such as water, vanadium, and aluminum/silicon have been added.

Residual Fuels. The 15 residual fuels RMA 10 to RML 55 are grouped according to viscosity, their numbering being identical with the viscosity at 100°C, rising in steps of 10 mm^2/s with one intermediate fuel of 15 mm^2/s. Each viscosity group contains either two or three fuels, allowing different concentrations of asphaltenes described by variations in, e.g., densities, carbon residues, or vanadium content. Fuel manufacturers' specifications use often the viscosity at 50°C and their fuel denotations thus differ from that used in ISO 8217. In the traditional marine business the viscosities at 50°C are still the most widely used viscosity parameters.

While initially the maximum density of the fuels was restricted to about 0.990 kg/m^3 to guarantee the separation of water by centrifuging, modern centrifuges can safely handle fuels with densities even above 1000 kg/m^3 and for that reason the maximum density of some of the more viscous fuels was raised to 1010 kg/m^3.

The flash point (safety aspect in storage tanks) of all residual fuels is identically specified to be ≥ 60°C. In theory the fuels should not produce ignitable air-fuel mixtures in the fuel tank atmosphere below this temperature.

Paraffinic wax formation at low temperature is controlled by specification of the pour point. The pour point is 3°C above the temperature at which the fuel will not flow under its own weight when tested by standard methods. The pour point is a guide for low temperature filterability/pumping limits and wax precipitation in storage tanks. (In automotive diesel fuel the cold filter plugging point, CFPP, a low-temperature filterability test, has shown better correlation with low-temperature fuel filterability performance). As residual fuel oils are heated anyway to reduce their viscosity, wax formation even with the high pour points of 30°C causes generally no problems. However, care must be taken that the fuel is not cooled and heated in cycles around the pour point, because under these conditions wax formation may increase. It should be noted that some paraffinic wax crystals once formed in residual fuel oils have melting temperatures approaching 100°C.

Ignition performance of residual fuels is not specified, because there is sufficient time in the low-speed – long-stroke engines to burn even highly aromatic/asphaltenic fuels with low ignition performance. It was also shown that engine and operating parameters (e.g., fuel viscosity when injected, charge temperature and pressure) can compensate fuel deficiencies in this aspect. When changing from one fuel to another, it may, however, be necessary to rank their ignition performance. It is well known that that the ignition performance of a diesel fuel decreases with increasing aromatic content. In the absence of both, a suitable and simple test method of ignition performance and of aromaticity of residual fuels, the Calculated Ignition Index (CII) and the Calculated Carbon Aromaticity Index (CCAI) have proven useful. Both methods use fuel viscosity and density as parameters. Increasing viscosity and density simultaneously increases the aromatic content and reduces the ignition performance. At a given density reducing viscosity and at a given viscosity increasing density increases the calculated aromatic content and reduces the ignition performance of the fuel. Details are described in an appendix of ISO 8217.

The carbon residue of marine diesel fuels is measured to estimate the coke forming and fouling tendency of a fuel. It is generally related to the asphaltene content, but there are considerable differences in this relationship between different fuels. The correlation of carbon residue with formation of deposits in marine engines is poor and that is not surprising because modern marine engines (specifically low- and medium-speed engines) are built to operate safely with fuels of high asphaltene contents and high carbon residue. However, the carbon residue has been retained in specifications to characterize the fuel.

The ash content of fuels is specified, because inorganic ash-forming matter, either of natural origin or originating from contamination, can form deposits in the turbocharger and in the engine.

Specifically deposits on the exhaust valve seats can lead to valve burning and finally failures. Whereas insoluble and water-soluble material (e.g., rust and sodium salts from sea water) are generally removed by fuel centrifuging, “natural” oil-soluble ash-forming impurities (mainly various metals and silicon), which tend to concentrate in the distillation residues, cannot be separated on board.

Water is the most common contaminant of marine fuels. Water content is often increased after delivery in the ships fuel tanks, either because of leakage into the tank or by tank breathing and condensation. Water bottoms in fuel tanks can lead to rust formation and to the growth of fungi in the water-oil phase. This may cause additional sludge in storage tanks. Fuel centrifuging separates most of the water before the fuel is supplied to the engine.

Sulfur is a natural constituent of crude oil. It is chemically bound to the hydrocarbon molecules and the sulfur concentration varies depending on the crude oil source, ranging from about 0.1% (e.g., Nigerian crude) to nearly 3% (e.g., Venezuelan crude). In distillation residues, used for the blending of residual fuels, the sulfur content is even higher. Heavier grades may contain a maximum of 5%. During combustion the sulfur is converted to sulfur oxides. Together with the water formed by combustion the sulfur oxides form sulfuric acids, which may cause severe corrosion of cylinder liners and piston rings. By the use of highly alkaline cylinder lubricants and higher engine temperatures to avoid condensation, the engine wear problem has been largely overcome, however, there remains the emission of sulfur oxides with the exhaust gases. The use of high-sulfur fuels is therefore restricted or penalized in some coastal areas.

Also vanadium is—like sulfur—a natural trace constituent of crude oil. During combustion vanadium together with sodium may produce in a series of complex reactions sticky and low-melting salts, which may adhere to exhaust valves and turbochargers, causing high temperature corrosion. Apart from restricting the concentration of vanadium in residual fuels there are operational (reduction of exhaust valve temperature) and metallurgical (specific alloys) solutions to avoid this type of corrosive wear.

Aluminum and silicon in fuel oils stem from disintegrated cat cracker catalyst. Of course these catalyst fines may only be present when cat cracked components are used as components. The catalyst fines may cause abrasive wear. Below the specified limit of 80 mg/kg they are safely reduced by centrifuging to uncritical levels.

The total sediment contains not only inorganic insolubles but, after aging of the fuel, also insoluble organic fuel components such as agglomerated asphaltenes. As mentioned above, residual fuels contain significant amounts of asphaltenes which are kept finely dispersed by aromatic diluents. Fuel stability tests are main criteria of quality control in refineries and it is unlikely that any stability problem producing asphaltenic sludges occur with fuels supplied from refineries according to in-house and international standards. However, if a fuel is blended with paraffinic diluents or fresh asphaltenes are produced by heating (e.g., with electrical heaters with too high skin temperature), it is possible that the asphaltene/aromatic ratio is altered and the asphaltenes may agglomerate adding considerably to sludge formation. Experience has shown that residual fuels with total sediments below the specified 0.1% can be handled without any problems with the standard centrifuging equipment.

Another important quality criterion, which is not specified, is the fuel’s specific energy. The specific energy can, however, be calculated sufficiently precisely from the inspection properties of the fuel, if known. For details see ISO 8217. Increased density (i.e., higher carbon/hydrogen ratio), sulfur and, of course, water and ash content reduces the specific energy on a weight basis (in MJ/kg). However, fuels belonging to the high density/viscosity categories have usually lower prices, their deficiency of lower specific energy content is therefore generally more than offset by the price.

[Top of Page]

5. Outlook

While for several decades fuel research and engine development was geared to reduce overall cost by improving efficiency and engine life with cheap fuels, the focus has changed to improved environmental acceptability. In the area of marine diesel emissions this is mainly the reduction of harmful exhaust gas constituents. Depending on the nature of these pollutants there are different options, i.e., changes in fuel composition, of the combustion process, or exhaust-gas aftertreatment. Different technologies are available and have partly been proven “on shore” such as selective catalytic reduction (SCR) exhaust-gas aftertreatment catalysts to reduce nitrogen oxides. Research in all three areas is ongoing, however cost and practicability on board are some of the limiting factors. International standards on emissions taking into account the technical potentials with acceptable lead times or penalties, as already enforced in sensitive areas like the Baltic sea, help to neutralize the cost factor and will lead to environmentally improved ship propulsion systems.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Chemical Plants: Performance Measurement of Processes

NASREDDIN DHAFR, School of Science and Technology, University of Teesside, Middlesbrough, UK
MUNIR AHMAD, School of Science and Technology, University of Teesside, Middlesbrough, UK

1. Introduction 1
2. Performance Measurement Practices 3
2.1. Performance Measurement Frameworks 3
2.2. Performance Measures and Key Performance Indicators 4
2.2.1. Corporate Performance Indicators 5
2.2.2. Strategic Key Performance Indicators 5
2.2.3. Operational Key Performance Indicators 6
2.3. Performance Targets for Chemical Process Plants 7
2.4. Performance Measurement Problem in Chemical Process Plants 8
2.5. Performance Objectives of Batch Chemical Plants 8
3. Operational Effectiveness Model 9
3.1. Structure of Operational Effectiveness Model 9
3.1.1. Step 1: Manufacturing Process Inputs 9
3.1.2. Step 2: Effectiveness Measures . 10
3.1.3. Step 3: The Actual Process 10
3.1.4. Step 4: Theoretical Determination of Performance Targets 10
3.1.4.1. Determination of Ideal Cycle Time 11
3.1.4.1.1. Prediction of Heat-Transfer Time 11
3.1.4.1.2. Chemical Reaction Cycle Time . 12
3.1.4.2. Defining Ideal Yield 12
3.1.4.3. Defining Ideal Production Quantity (Economic Batch Quantity) 12
3.1.5. Step 5: Import of External Benchmarking Data 13
3.1.6. Step 6: Comparisons and Identification of Gaps 13
3.1.7. Step 7: Definition and Application of Improvement 14
3.2. Example of Practical Application 14
4. Conclusion 15
5. References 15

Performance measurement is a competitive tool widely used by manufacturing and service organizations to evaluate the output of business processes. Its fundamental purpose is to quantify the potential for improvement and construct plans for further development. This article aims to increase the understanding of performance measurement in chemical plants. Existing performance measurement practices are reviewed, and their roles in the improvement of business and operational results of chemical process plants are defined. A developed performance measurement model, called the "operational effectiveness model", is also introduced. It comprises a sequence of several performance measurement steps which ensure that chemical batch plants operate efficiently and effectively.

1. Introduction

Performance measurement is a competitive tool and it has been widely used by manufacturing and service organizations to evaluate the output of business processes. The fundamental purpose of performance measurement is to quantify the potential for improvement and to construct plans for further development [1 – 3]. The recognition of the importance of performance measurement in manufacturing and the need for linking manufacturing strategy and performance measurement can be traced back to a specific article [4]. Since that article, the number of strategic-oriented performance measurement tools has increased, but selecting the right one is a real chal-

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.k06_k01

lenge. This selection difficulty is caused by the variation between different manufacturing processes, a dissimilarity of business objectives, and the depth of understanding of specific business requirements.

Performance measurement is used to evaluate, control, and improve a company's production process, to ensure the achievement of its goals and objectives. If results differ from objectives, an organization can analyze the gaps in performance and make adjustments. Performance measurement is also used to compare the performance of different companies, plants, and individuals. In addition, performance measurement provides a means of capturing performance data, which can be used to inform decision making [5].

It is contended [6] that measurement plays a crucial role in translating business strategy into results. In a survey of top U.S. companies [6], it was found that companies with effective measurement tend to anticipate the future and are likely to remain competitive in a radically changing environment. Performance measurement yields many benefits for an organization. One benefit is that it provides a structured approach to focus on a company's vision, strategy, goals, and performance. Another benefit is that it provides a mechanism to report program performance to high-level management. Besides, performance measurement can help companies [7]:

- Track progress in achieving desired performance goals.
- Control performance within predetermined boundaries.
- Identify problem areas and possible problem causes.
- Show where improvement needs to be made.
- Ensure that decisions are based on fact, not on emotion.
- Identify whether suppliers are meeting requirements.
- Compare and benchmark competitors' performance with their own.
- Set future targets.
- Drive change.

World-class manufacturers are at the top of their industries as a result of their consistent delivery of superior performance [1, 8, 9]. This superior performance is achieved by continual assessment and improvement of business processes, application of the proven best solutions to improve the performance of key business areas, and continual measurement of financial and operational performances [1, 10, 11].

The application of performance measurement practices produced a tool called "benchmarking". Rank Xerox defined this [12] as "a continuous, systematic process of evaluating companies recognized as industry leaders, to determine business and work processes that represent best practices, and establish rational performance goals. There are four types of benchmarking [12]:

- *Internal benchmarking* applies more to large organizations that have various business units. It is about comparing internal operations from one sister company to another.
- *Competitive benchmarking* specifically compares competitor to competitor, using the product or function of interest. This goes beyond traditional product or service engineering, but it is important to learn and understand about competitors' methods, their processes, innovation, strategies, markets, etc.
- *Functional benchmarking* is the comparison of similar functions within the same broad industry, using wide industry leaders as partners.
- *Generic benchmarking* involves establishing the comparison of business functions or processes that are the same, regardless of type of industry. Generic benchmarking is a long-term challenge and reflects a total change in the culture. It is usually found in learning organizations.

The fact that process industries are less familiar with measuring and benchmarking their own process manufacturing performance has been discussed in the literature [1]. The argument often used for why this unfamiliarity occurs is that the process industries are different; therefore, the implication is that performance measurement and benchmarking techniques do not apply [1]. These techniques, however, have been developed and successfully applied in other manufacturing industries such as electronics, retail, and automotive.

This article aims to increase the understanding of performance measurement in chemical plants. The objective is to assist these plants in their journey to attain effectiveness in op-

erational performance. In this article, the existing performance measurement practices are reviewed, to define their roles in the improvement of business and operational results of chemical process plants. A developed performance measurement model, called the "operational effectiveness model", is also introduced. The model comprises a sequence of several performance-measurement steps that are put together to ensure that operations at chemical batch plants are operating efficiently and effectively.

2. Performance Measurement Practices

Companies in manufacturing or service normally define objectives from their manufacturing mission and vision [13], then define and implement their manufacturing strategies to attain these objectives [14]. The objective is the statement of a general condition that needs to be achieved and the strategy was defined as "the art of devising or employing plans toward a goal" [15]. The strategy to attain the performance objectives often requires the selection of some appropriate tools that can drive the actual performance of the process towards the levels of performance targets (benchmarks) to ensure that the process is operating effectively. The performance measurement system is used to assess the output results of the adopted strategies, with respect to customer expectations and strategic objectives [4, 12, 16 – 18]. As suggested [19], however, such measurements will be more useful if they can also serve as a diagnostic aid to better understand the reasons why a company is performing well or poorly; therefore, a well-designed system of performance measurement must incorporate the ability to both monitor and diagnose performance. The performance measurement system is often constructed as in the performance measurement system design process [13]. The performance measurement system, however, is a constructed framework comprising different elements, such as how performance would be measured, which area or activity in the process should be measured, and which measures and which type of performance targets would be used. It is stated [20] that good systems include the need to link operations to strategic goals, integrate financial and nonfinancial information, measure what is important to customers, motivate operations to exceed customer expectations, and identify and eliminate waste.

It is said [21] that an effective performance measurement system should consist of not only the traditional financial and cost-accounting criteria used by upper management, but also tactical-performance criteria that are used to assess the firm's current level of competitiveness and direct its efforts to attaining a desired competitive position. The roles of performance measurement can be categorized into three aspects:

1) It provides the company with a method to assess its current competitive position with respect to its competitors and the demands of the market and identify avenues for improvement.
2) It monitors the company's progress in moving towards its strategic objectives [13].
3) It highlights where improvements have been made and which areas still need to be improved [22].

In Sections 2.1, 2.2, and 2.3, the performance measurement frameworks, performance measures, and target setting methods are briefly described for the purpose of highlighting requirements for measuring results of adopted strategies and attaining the objectives of chemical plants.

2.1. Performance Measurement Frameworks

Several performance measurement frameworks were developed to assist practitioners and researchers in measuring the performance of organizations in the manufacturing or service sectors. It was stated [23], that the first development of the "tableau de bord" performance measurement system was carried out in the early 1900s, by a group of French engineers. In the last three decades, however, several performance measurement frameworks were introduced, such as the strategic measurement analysis and reporting technique (SMART) [20]; the balance scorecard [24]; the performance prism [25]; and the hierarchical performance measurement system [26]. These performance measurement frameworks proved to be effective tools for measuring the performance of organizations and also

assisted businesses to shift from focusing on traditional financial results, to focusing on new dimensions related to employees, customers, and other stakeholders. These frameworks are found to be useful for application at different levels in the organization, as they provide management with a complete set of performance data to inform decision making. They are, however, lacking the details required by engineers on how to define targets, tackle causes of poor performance, define improvement areas, and implement best practice.

A methodology was presented on how to measure and benchmark the performance of process manufacturing plants [1]. Several categories of performance measures were defined for this purpose. These categories of performance measures are customer services; reliable assets; operational excellence; motivated people; and safety, health, and environment. This methodology has also recommended the use of fewer focused measures, to make the process more effective. Another approach, called the "manufacturing enterprise performance measurement system", was introduced [11]. It is a top-down approach for the manufacturing organization, in which the organization's objectives, strategies, and performance measures are identified at different levels of the organization. By applying the methodology of benchmarking performance of the process plants in the manufacturing enterprise performance measurement system, it was found to be useful for measuring the levels of plant, department, or business in the organization, but bottom-line operational levels are difficult to measure by this methodology, as this level requires specific techniques to define performance targets and demonstrate performance results. The operational level at the plants requires a framework that can handle the technical details of the operations and actions of the bottom line process and identify proper performance targets, based on the capability of the process.

2.2. Performance Measures and Key Performance Indicators

Manufacturers are required to provide their products to users at the expected time, with reasonable prices, and meeting the required quality [27]. As a result, several performance measures are used in practice and research, to give an indication of the performance of process plants, as well as plants from other manufacturing sectors. Traditionally, the measures used to evaluate a company's performance have been based primarily on conventional accounting systems. Examples of these measures include earnings per share, inventory turns, return on investment, purchase price variance, labor efficiency, and machine utilization. Such measures are focused on local measures of efficiency and productivity [5]. Among them, productivity and cost are the most frequently used measure of manufacturing performance. Manufacturing companies monitor employee productivity, facility output, and direct and overhead costs, for a variety of time frames. This is because only costs are available and easy to measure [28]. The purpose of financial accounting measurement [29] is to report the company's activities to interested parties outside the company, whereas a company's cost accounting measure is supposed to provide information useful to managers' planning and control decisions. Financial accounting measures, in general, would provide managers with knowledge of the system performance that reflects the results of past decisions, but not the actionable steps needed for surviving in today's competitive environment.

As a result of the limitations of the traditional performance measures, attempts have been made to develop a set of better-balanced measures that would provide management and operators with the on-time information that is necessary for daily decision making [30]. Such measures should be flexible, primarily nonfinancial, and able to be changed as needed [31]. A major benefit of nonfinancial measures is that they can usually be reported on a timelier basis than financial measures. As a consequence, if a nonfinancial measure indicates poor performance, action can often be taken before negative financial consequences occur. In addition, nonfinancial measures are easy to understand and are always available. This provides instant feedback regarding the performance of the company [32].

The key performance indicator (KPI) is one type of performance measures. The KPI is a number, or value, which can be compared against an internal or external target performance benchmark, to give an indication of per-

formance. That value can relate to data collected or calculated from any process or activity [33, 34]. The KPIs are often classified as strategic-in-nature KPIs or operational-in-nature KPIs. Operational KPIs are used by managers to plan and control programs at operational level, while strategic KPIs provide guidance to both managers and policymakers who have to make decisions from a more global perspective. Generally, KPIs should be selected according to how they relate to the strategic aims of the organization and how they can support the corporate-level performance indicators. This means that KPIs, whether they are operational or strategic in nature, must tie in to strategy.

In general, in chemical batch plants, there is still a requirement for operational measures that can support achieving the required level of performance in the strategic KPIs.

2.2.1. Corporate Performance Indicators

Corporate performance indicators are related to achievement of corporate objectives; some of these performance indicators are sales growth, customer loyalty, assessment of stakeholders' satisfaction, percentage market share, profitability measures, return on investment (ROI), return on capital employed (ROCE), and return on net asset (RONA).

Customer loyalty is a qualitative performance measure, ranging from "never buying from anywhere else because the experience is so satisfying" to "returning, because there is no alternative available".

Sales growth can be calculated from self-reported data (Eq. 1).

$$\text{Sales growth} = \frac{(\text{Total sales this year} - \text{Total sales last year})}{\text{Total sales last year}} \quad (1)$$

Sales figures do not necessarily indicate how a firm is performing relative to its competitors. The firm's performance relative to competitors can be measured by the proportion of the market that the firm is able to capture. This proportion is often referred to as the market share (Eq. 2).

$$\text{Market share} = \frac{\text{Firm's sales}}{\text{Total market sales}} \quad (2)$$

Firms that have high customer loyalty usually have a high market share in their defined markets. Market share is often associated with profitability; thus, many organizations seek to increase their sales relative to competitors.

The return on investment (ROI; Eq. 3) is a measure used to determine whether a proposed investment is required and how well it will repay the investor.

$$\text{ROI} = \frac{\text{Profit}}{\text{Investment}} \quad (3)$$

Return on capital employed (ROCE; Eq. 4) is probably the most popular ratio for measuring general management performance in relation to the capital invested in the business. ROCE defines capital invested in the business as total assets less current liabilities.

$$\text{ROCE} = \frac{\text{Net profit before interest and taxes}}{\text{Total capital employed}} \times 100 \quad (4)$$

Return on net assets (RONA; Eq. 5) is another measure of financial performance. The higher the return, the better the profit performance for the company.

$$\text{RONA} = \frac{\text{Net income}}{\text{Fixed assets} + \text{Net working capital}} \quad (5)$$

With regard to shareholder value, the key driver is profit, or the expectation of profit.

2.2.2. Strategic Key Performance Indicators

Several key performance indicators are often used to measure the strategic performance of manufacturing processes. The "adherence to production plan" (Eq. 6) is the main key performance indicator for the effectiveness of manufacturing processes to meet production planning. It is the percentage of planned activities on the sales and operations plan (S& OP) that are completed as planned in a planning period [1]. Deviations from the plan are sometimes caused by equipment failures or by poor demand forecasting. This key performance indicator indicates the opportunities for improving the demand management process [1].

$$\text{Adherence to production plan} = \left(1 - \frac{\text{Actual production} - \text{Forecast production}}{\text{Forecast production}}\right) \times 100 \quad (6)$$

Overall equipment effectiveness (OEE; Eq. 7) was developed in the 1980s [35]. This key performance indicator measures the effectiveness of equipment's deployment to perform work over a period of time:

$$\text{OEE} = \text{Performance rate} \times \text{Quality rate} \times \text{Availability} \quad (7)$$

The production losses, as defined by the OEE, are related to equipment breakdown, setup, minor stoppage, reduced speed, scrap rate, and startup. Many batch and continuous plants in the chemical industry have used OEE to measure the performance effectiveness of their operations and have used the results to tackle the causes of low performance and optimize outputs. The main criticism of OEE, however, is that all the measured losses are equipment-related losses.

The process capability (CpK) [1], is the flexibility and quality key performance indicator. It measures the encountered variability of process results without effect from any special cause (Eq. 8)

$$\text{CpK} = \frac{\text{Lower of (upperspec} - \text{average)}}{3\sigma} \quad (8)$$

where upperspec is the upper level of allowable deviation from the average, and σ the standard deviation.

The on time in full (OTIF; Eq. 9) is the supply chain's main key performance indicator. It measures the percentage of all deliveries that are delivered to the final customer on the mutually agreed date and location, defect-free in all respects. This includes the product quality, order size, packing, and support paperwork [1].

$$\text{OTIF} = \frac{\text{Number of defect-free on-time shipments}}{\text{Total number of all shipments}} \quad (9)$$

The stock turn (Eq. 10) is a financial key performance indicator. It is an important KPI for internal issues and also needs to be as high as possible, to ensure that the capital is not tied up on inventory.

$$\text{Stock turn} = \frac{\text{Cost of sales}}{\text{Values of all materials in stock}} \quad (10)$$

The value of all materials in stock comprises the value of raw materials, work in progress, finished goods, and engineering items. The reasons for holding stock are many and varied [1]. It could be strategic stock in anticipation of some process or plant unreliability, or a change in the world supply market; financial reasons, where any significant price rises are anticipated and the products have been bought ahead of this; stock may be held due to process unreliability, in anticipation of planned shutdowns, or in consignment form to ensure that the customer is never let down. Hence, stock turn may be used as a measure of customer service or operational excellence.

The value added per manufacturing employee (Eq. 11) is also a financial key performance indicator. It is the value of sales turnover, from which all goods and services purchased are subtracted, divided by the number of direct employees working in manufacturing [1].

$$\text{Value added per manufacturing employee} = \frac{\text{Value of sales} - \text{all costs of the production}}{\text{Number of manufacturing employees}} \quad (11)$$

The KPIs described above are considered as a strategic level set of measures [11]. The sequence of these KPIs, however, gives a complete set of strategic performance measurements, related to different strategic dimensions. The appropriate combination of measures is determined by the nature of the actual industry. A bulk chemical industry, for example, may focus more on the adherence to sales and operational planning (S& OP), whereas a consumer product industry is encouraged by its final customers to pay particular attention to issues such as OTIF and customer complaints. In general, these KPIs support the corporate performance indicators at the corporate level of the organization, such as return on investment, return on assets, market share, customer loyalty, stakeholder satisfaction, and so on.

2.2.3. Operational Key Performance Indicators

Operational KPIs are often selected with respect to their relevance to the operational objectives and the strategic KPIs. Some of the commonly used operational measures are as follows:

- *Manufacturing cycle time*: the total time that the process takes, from start to finish, to produce the required product. This time normally comprises the time required for cleaning, loading, processing, and emptying the equipment.

- *Value of stock in warehouse*: all the money the system invests in purchasing things the system intends to sell.
- *Cost reduction*: KPIs in this category measure the achieved cost reduction through an improvement programme.
- *Amount of wasted materials*: material losses from processing, quality control inspections, handling, and transfer of materials.
- *Quality measures*: these measures comprise right-first-time performance, defects per unit of production, and detailed pareto diagrams, etc.
- *Equipment uptime*: uptime is defined as the time a process operates at its standard maximum proven rate while making a defect-free product. The ratio of uptime to the total available time is often calculated by Equation 12.

$$\text{Uptime}=\frac{\text{Valuable operating time}}{\text{Valuable operating time}+\text{Total losses}} \quad (12)$$

- *Equipment downtime*: downtime is defined as the period when the equipment is not available to fulfill its intended role. Downtime can consist of planned activities, such as scheduled maintenance, repair, and overhaul, and unplanned activity as a result of a failure or accidental damage.
- *Resource utilization* is defined as the proportion of time that a resource is used.
- *Final product yield*: for a description of this measure, see Section 3.1.4.2.
- *Production capacity* is defined as the maximum rate of output that a production unit is able to produce under a given set of operating conditions.

The basis for calculating operational performance measures, however, differs from one sector to another, from plant to plant, or even from one operation to another operation within the same plant. The manufacturing cycle time is an example, as, in some cases it may be calculated by simple, straightforward equations, or in other cases, may require complicated calculations.

2.3. Performance Targets for Chemical Process Plants

The target is the level of result that needs to be achieved. The purpose of performance measurement is therefore to compare actual performance results with an established set of performance targets [36]. There is still little attention paid to this area, in terms of definition and explanation of effective methods, to define the performance targets for performance measures. It was stated [37] that many companies have a problem when performing benchmarking activities. This problem is caused by a difficulty in retrieving quantitative data relevant to world-class or competitors' performances. It was also highlighted [37] that there is little work being done to address this problem or, at least, to propose a dynamic technique that can help manufacturing organizations in this area. When data for benchmarking are not available, the organization is advised to establish targets the best it can [37]. Organizations have always set performance targets and used measurement systems to track the achievement of targets [17, 18, 29, 38]. The targets were classed in four categories [29]:

- *Competitive benchmarking*: in this category, a detailed analysis is made of competitors' operations and an attempt is made to emulate the best aspects of each competitive company.
- *Price targeting*: this category includes the application of cost-control techniques in detail, throughout the initiation and design phases of a new product. This programme results in a detailed breakdown of cost and quality targets for each department of the company.
- *The half-life concept*: this category includes an analysis of the rate of improvement that can be expected for different kinds of continuous improvement tasks. This system provides a more objective method of setting goals for quality improvement programs.
- *Common sense and a world-class manufacturing mind-set*: this category includes the application of realistic and motivating targets for companies involved in the introduction of world-class manufacturing techniques and continuous improvement.

Four methods that can be used to set or generate performance targets were also defined [17, 18]:

- *Historical targets*: this method is often used, owing to its simplicity in implementation [39]. The best historical performance of an activity would be defined, then used as a performance target.

- *External targets*: by using this method, the world-class performance of best-in-class operations is collected to use for benchmarking purposes. Table 1 lists some superior performances of process industries' world-class plants that are often used as performance targets. Importing external benchmarks from competitors' performances, however, is often an essential requirement for competitiveness in any business.
- *Internal targets*: these targets are based on comparisons of the performance of internal processes. In this method, the performance of the best performing process in the company is used as a benchmark. This method is myopic, however, in the sense that it looks only within the organization and may limit the company's performance relative to its external environment [39].
- *Theoretical targets*: in this method, the targets are determined by the use of theoretical calculations, or by simulation analysis. Setting performance targets on a theoretical basis is exceptionally useful, as it defines the performance capability and limitations of the process. A considerable disadvantage of this method relates to the difficulty and time needed to define it. The theoretical targets of the process would enable effective comparisons of process performance to benchmarks from external sources; they would help in assessing the capabilities of operations and whether or not they are able to perform similar to best-in-class operations.

Table 1. World-class performance of process plants [1]

No.	Key performance indicator	World-class performance
1	Adherence to production plan	> 99 %
2	Overall equipment effectiveness (OEE)	> 95 % for continuous plants> 85 % for batch plants
3	Process capability (CpK)	> 2
4	On time in full (OTIF)	> 99.9 %
5	Stock turn	> 25
6	Value added per manufacturing employee	$400 000
7	Training days per employee	12 d
8	Absenteeism	< 1 %

2.4. Performance Measurement Problem in Chemical Process Plants

The continuous mode of operating process plants has, to some extent, fewer management problems at operational levels. The equipment in this mode is designed and scaled to produce bulk chemicals with a defined rate of production. In addition, most material-handling and -moving steps are automated. In this case, defining the ideal performance target is not a difficult task; therefore, applying the operational, or strategic KPIs in the plant is rather simple, when compared to the batch process plants.

In the case of batch mode, performance measurement often encounters some problems. In this mode of production, for example, flexibility in the product formulation and production rate causes uncertainties when selecting an optimum set of operation variables. The manufacture of different products in the same production equipment will also require frequent changeover and equipment setups. Consequently, cleaning at the changeover between two different products causes the loss of a significant amount of materials. A further problem is that the flexibility achieved by batch processing, which enables the production of larger or smaller quantities, also raises the question of the cost effectiveness of the selected size of the batch.

Despite the fundamental improvements made by applying manufacturing practices to the performance of manufacturing plants, there is still scope for improvement in chemical batch processes, as new tools are needed for defining a complete set of performance measurement requirements. Therefore, research, in such areas as defining performance objectives, defining manufacturing strategies, and defining elements of performance measurement system, is fundamentally important to chemical batch plants.

2.5. Performance Objectives of Batch Chemical Plants

The manufacturing objectives of the batch chemical plants were defined by a review of relevant research literature and on informal interviews with some key individuals in batch chemical plants. The conclusion of the study was that

the minimum production cycle time, maximum yield of final product, and minimum total production costs are the main three objectives of the batch chemical processes [18, 40, 41].

There are several benefits of cycle-time reduction, some being minimum waste in time, cost reduction, increased flexibility, efficient employees, and less time exposed to risk while operating hazardous processes.

The main benefit of increasing the yield of the final product is the utilization of raw materials. Raw materials make up the largest portion of manufacturing costs; therefore, when they are used inefficiently, waste will be created, which will lower sales revenue and create an environmental problem [42].

Economic batch quantity is an essential requirement to achieve the minimum cost of manufacturing. Capacity decisions are vital to the manufacturing plants, as they affect a number of dimensions in the business, such as delivery date, product prices, and warehouse requirements. Costs associated with production setups, resource utilization, and inventory holding are all related to the batch quantity and contribute to total manufacturing costs. The link between stock and the financial performance of organizations is strong [43]. The organizational objective "maximizing return on assets" is a good example of this link. Another advantage of manufacturing at economic batch quantity is the tuning of the output of manufacturing either to meet only the market demand, or to produce for demand and inventory. Therefore, when the manufacturing process is optimized for minimum cycle time and maximum yield, then the economic batch quantity will ensure that the plant will not overproduce to meet the objective of those two factors.

In Chapter 3, a detailed model is introduced, showing how to measure performance of the chemical batch plants with respect to the three performance objectives.

3. Operational Effectiveness Model

This chapter of the article introduces the operational effectiveness model as a performance measurement tool for chemical batch plants, as shown in Figure 1. The proposed model is developed from the framework introduced in [18]. The steps of this model are discussed in terms of how they are developed for the specific requirement of manufacturing in chemical batch plants. This high-impact model is planned to be capable of developing tools for industry, to redefine their operational effectiveness requirements.

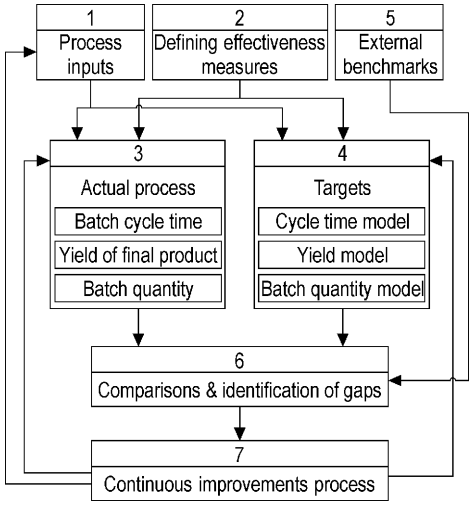

Figure 1. The operational effectiveness model for measuring performance of a batch chemical process.

3.1. Structure of Operational Effectiveness Model

The model begins with the development of an understanding of manufacturing process inputs and the fundamental requirements that are placed on them. In another concurrent step, the pertinent performance measures must be identified, which in this case is the introduction of a new key performance indicator, called the "batch effectiveness factor" and its calculation requirements. Next, the model's actual manufacturing process is defined concurrently, and ideal performance targets are determined. Performing comparisons, performance analysis, and a signposting route for improvement steps, are also included in the model.

3.1.1. Step 1: Manufacturing Process Inputs

The process inputs are key elements in the manufacturing process that will influence its perfor-

mance. The process inputs must be selected to meet certain standards, such as the specification of raw materials, skills of operators, reliable equipment, adequate scheduling, and the right sources for process utilities. All these requirements are process inputs, or resources, that are required in the majority of manufacturing processes. These inputs would interact with each other in performing the required product with the required quality, time, and amount. Information from external stakeholders must be collected and analyzed for any requirements that may be included in the process. Sources of external requirements could include downstream customers, governmental organizations, or labor unions. Customer requirements placed on the manufacturing process will often include factors related to product quality, price, and delivery date.

3.1.2. Step 2: Effectiveness Measures

The selected performance measures in this model are: effectiveness of cycle time, specific yield, and effectiveness of capacity. The three measures are single performance measures for defining performance of the batch process. Each of them measures the process with regard to one performance objective. The performance effectiveness of the process in relation to each objective is calculated by comparing actual performance of the process to the performance targets (Eq. 13).

$$\text{Performance effectiveness} = \frac{\text{Actual performance value}}{\text{Target performance value}} \times 100 \quad (13)$$

The results of the three performance effectiveness measures are then combined in one factor, called the *batch effectiveness factor*, which gives an indication of the effectiveness of the batch process, in relation to the three objectives, as shown in Equation 14.

$$\text{Batch effectiveness factor} = \text{Cycle time effectiveness} \times \text{Specific yield} \times \text{Capacity effectiveness} \quad (14)$$

The development of the batch effectiveness factor aims at unifying the results of performance effectiveness with respect to the three objectives in one factor. Performance assessors therefore can determine whether the process is achieving thorough effectiveness without being misled by the value of effectiveness in meeting one objective.

3.1.3. Step 3: The Actual Process

The process is the main part of manufacturing; therefore, the process step is set as the core of the operational effectiveness model. In the process, the actual inputs and resources are transformed to create the required product. The utilization of resources is often a crucial issue for many manufacturing enterprises. In an ideal manufacturing operation, the processes would operate at full utilization of all resources, for example, full utilization of manpower and equipment time, full utilization of capacity, zero waste from handling and processing of raw materials, and no scrap from the process utilities. This situation, however, is rare in real operations. The difference between the ideal and the real situation is due to losses [17, 18]. Based on the adequacy of inputs, resources, tools, and the extent of utilizing inputs or resources, production of the required product will either meet or deviate from the required performance targets, and hence the performance objectives. Any output of the process is expected to be a product, byproduct, or waste. Increasing the value in the finished product or minimizing the amount of unwanted byproduct, material waste, wasted time, and manpower requirements would result in an increase in the desirable outputs [44].

3.1.4. Step 4: Theoretical Determination of Performance Targets

The targets used in the calculation of effectiveness measures, as defined by Equation 13, will become the subject for the theoretical definition of the ideal performance of batch processes. Mathematical calculations, taken from the science of each selected operation or action in the process, will be used. The theoretical model for generating a performance target would represent the behavior of the process as it manufactures a batch from start to finish. Such a view would provide the necessary information for the

creation of a simulation model of a batch process, which would include routing information, process times, and, among others, labor requirements. The actual input data will be used in the calculation of the theoretical targets. The generated ideal results will be tested for their practicality; if the results are satisfactorily accepted, then they can be chosen to become the process performance target. The approach of theoretical performance targets is an attempt to provide an answer to the fundamental question "what must the process do?". There is an understanding that the route towards the achievement of world-class level of performance is through the full knowledge of the level of results that can be obtained from the process.

In some cases, the model aims to investigate the heat-transfer steps in chemical plants, to predict the time needed for achieving the required heating or cooling. In chemical engineering literature, a procedure exists for prediction of heating or cooling requirement for the chemical batch process [45 – 47]. Another aim is to study chemical reactions in the chemical process and develop kinetic models for them. The variables affecting the rate of the chemical reaction require investigation, using laboratory and plant based data, with the aim of achieving maximum yield of the final product in a minimum cycle time. In the literature, there are also procedures to predict the time for materials mixing, separation, crystallization, and transfer. Generally, the determination of theoretical targets, intended for the operational effectiveness model, will include procedural examples of how to define the cycle time for selected operations in the process, defining the yield of the final product and defining the economic batch quantity.

3.1.4.1. Determination of Ideal Cycle Time

The measurement of cycle time is a primary feature of performance measurement for world-class manufacturers [29]. Ideal cycle time is the minimum time that a process takes from beginning to end, ideally without shutdowns or production-speed slowdowns. This ideal cycle time can be supplied in the nameplate by equipment manufacturers, derived from calculations, or obtained from a simulation of the process [18]. Some process operations and actions need theoretical investigations to assess the possibility of reducing their time and use of the ideal results as performance targets. These operations and actions include, for example, heat transfer, chemical reaction, material transfer, mixing, and separation.

3.1.4.1.1. Prediction of Heat-Transfer Time

The measurement and control of energy use within processes are interrelated, essential aspects of good energy management (→ Energy Management in Chemical Industry). The heat-transfer steps in the process require detailed theoretical study to predict the time frame for achieving the required heating or cooling. The prediction of heat-transfer time is required at the preliminary stages of the process equipment design, to assist the estimation of total process time and the required amount of energy to generate heating or cooling. Mathematically, for the chemical batch process, there is an existing procedure, related to the prediction of heating or cooling requirements [45 – 47]. Equation 15 is used to calculate the quantity of transferred heat Q from or into the system [48 – 50]. This equation is derived from Fourier's Law [51] and, when used for the batch reactor, the only variables are overall heat transfer coefficient U_o and the logarithmic mean temperature difference ΔT_{LMTD}, as the heat-transfer area A in the batch reactor is constant. The main drivers of improvement in the time needed for heating or cooling are therefore U_o and ΔT_{LMTD}. A higher value of U_o will decrease the heat-transfer time; therefore, a maximum value of U_o is needed for the objective minimum cycle time.

$$Q = U_o \, A \, \Delta T_{LMTD} \tag{15}$$

Properties of the equipment's material of construction, physical and chemical properties of the process and service fluids, and the operational conditions are the main factors affecting the value of U_o. Choosing raw materials with the right physical and chemical properties will assist in operating the chemical plant efficiently. The main physical properties affecting the performance of heat-transfer operations are viscosity, specific heat capacity, thermal conductivity, density, and specific gravity. Operating conditions such as flow rate of heat-transfer fluid and

temperature difference between the heat-transfer fluid and process fluid also play important roles in determining the value of U_o.

3.1.4.1.2. Chemical Reaction Cycle Time

The reaction rate is dependent on the reaction order. In this situation, the rate of a first-order reaction is directly proportional to the concentration of a single reactant [52, 53], and that of a second-order reaction increases as the square of that concentration [52]. In batch chemical processes, kinetics of the chemical reaction are needed in order to specify the type of reactor to be used (→ Process Development) and estimate the required time for the process, for use as a parameter in the design of plant equipment and also to estimate the profitability of the process. The variables affecting the rate of chemical reaction are temperature, concentration of reactants, surface area of reactants, pressure, presence of a catalyst, phase of reactants, presence of inhibitors, and equipment configuration [52, 53]. In the literature, the effect of such issues on increasing the rate of the chemical reaction has been discussed intensively. The optimization of the rate of chemical reaction is based on those issues and, in many cases, would result in a significant improvement of reaction time. Many research efforts were carried out to minimize the chemical reaction time for different processes. Generally, kinetics of the chemical reaction are critical to estimating the required time for the process, but detailed kinetic models, or mathematical optimization methods, are still seldom utilized in the industry to determine the optimal operation policy. Instead, a simple trial-and-error method and subsequent product quality check is preferred when deciding on the best operating policy [54].

3.1.4.2. Defining Ideal Yield

Manufacturers aim to obtain a high yield from each manufactured batch. The target is to achieve the highest potential transformation of raw materials into a desirable finished product. The existing definition of yield in the field of chemical engineering has limited improvement practice to the chemical reaction, or the transformational steps. Different definitions of yield and selectivity are indicated [55], and when reading literature on chemical reactions it is advisable to check carefully to ascertain the definition intended by the author. In the definition of yield [56], care must be taken to allow for the stoichiometry of the reaction. A decision must also be made whether or the yield should be calculated relative to the amount of reagent introduced into the system, or relative to the amount of reagent consumed. Work on the operational effectiveness model will relate to the yield of the whole process, in order to facilitate the definition of certain areas for improvement and to increase the desirable outputs. The effectiveness of the manufacturing process can be determined by calculation of the specific yield (Eq. 18).

The actual yield, theoretical yield, and specific yield can be defined [1, 41, 57, 58] as follows:

Theoretical yield y_{At} is the ratio of maximum potential amount of desired product to the ratio of all material inputs (Eq. 16).

$$y_{\mathrm{At}} = \frac{\text{Maximum potential amount of desired product}}{\text{Total amount of raw materials fed to the process}} \quad (16)$$

Actual yield y_{A} is the ratio of actual amount of produced desired product to the total amount of all material inputs (Eq. 17).

$$y_{\mathrm{A}} = \frac{\text{Actual amount of desired product}}{\text{Total amount of raw materials fed to the process}} \quad (17)$$

The specific yield y_{Asp} is the ratio of actual yield to theoretical yield (Eq. 18).

$$y_{\mathrm{Asp}} = \frac{\text{Actual yield}}{\text{Theoretical yield}} = \frac{y_{\mathrm{A}}}{y_{\mathrm{At}}} \quad (18)$$

The theoretical yield is a useful performance target for the actual process to aim at. The specific yield would then approach or reach 100 % — a value which is extremely difficult, if not impossible, to achieve [41], as the maximum amount of product is rarely produced. Product may be lost in material transfer, as a byproduct, or by inadequate operating procedures.

3.1.4.3. Defining Ideal Production Quantity (Economic Batch Quantity)

Manufacture at economic batch quantity is a strategy that is often used to attain the objective of minimum manufacturing costs [42, 59 – 61].

Measuring economic batch quantity is not new in science or practice, but it is rarely seen in chemical plants. Chemical batch plants produce to full equipment capacity for the following reasons:

1) Full utilization of equipment capacity and resources (Process Development)[55].
2) Safety inventory of finished products [1].
3) Concerns about the life cycle of liquid chemicals as raw materials.

With the emergence of world-class manufacturing practices and, in particular, the just-in-time (JIT) philosophy, the concept of full capacity utilization has been under intensive investigation, in order to compare savings from the utilization of resources during the production process and inventory costs. Following this, the concept of "economic batch quantity" was introduced into production planning and scheduling. The economic batch quantity (EBQ) is the quantity of production that operates at minimum total manufacturing costs (Eq. 19) [62]

$$\mathrm{EBQ}=\left(\frac{2C_{\mathrm{o}}\,D}{C_{\mathrm{H}}}\right)^{1/2} \tag{19}$$

where C_{o} represents batch-quantity-sensitive production costs, C_{H} inventory holding costs, and D inventory consumption rate.

The batch-quantity-sensitive production costs comprise the setup costs and costs of resources that are sensitive to batch quantity. The cost of holding inventory is known as the carrying cost and consists of warehousing, insurance, funds required, spoilage, and obsolescence.

Alternatively, the economic batch quantity can be determined as shown in Figure 2; in which the point for minimum total cost identifies the economic batch quantity.

Equation 19 and Figure 2 are explained in some detail in several textbooks [42, 59 – 61].

3.1.5. Step 5: Import of External Benchmarking Data

Performance of best-in-class operations is difficult to import at the batch level in chemical batch plants [11, 17, 18, 63]. This, however, should not delay efforts to collect the external benchmarking data. A comparison with other internal processes, or with other plants that have the same process, is also a useful option. A step in the operational effectiveness model has therefore been created to allow external benchmarks to play a role.

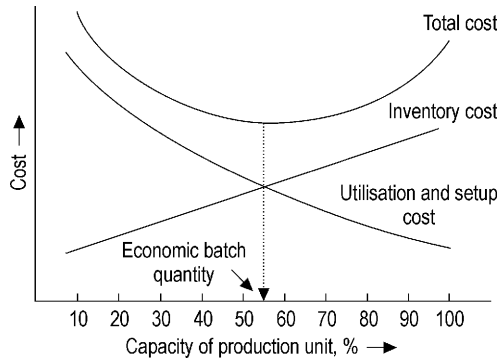

Figure 2. Defining economic batch sizes

Generally, care should be taken when using theoretical targets, as they are not enough in themselves; the competition may have deployed better equipment, technologies, or innovations to generate high-performance results. It is therefore advisable to use different target-setting methods in conjunction, as none of these methods are adequate, individually, for determining performance targets.

3.1.6. Step 6: Comparisons and Identification of Gaps

In Steps 3 and 4 of the operational effectiveness model, the actual and ideal performances of the manufacturing process were defined. The objective of performance targets is to compare the actual performance to them. In this case, the actual and optimum values of performance are defined, to give an indication of the process effectiveness, that is, close the actual performance is to its target. Each actual performance measured in Step 3 in the model is compared to the theoretical performance target determined by Step 4 in the model, to determine the effectiveness measures using Equation 13. At this stage in the operational effectiveness model, the batch effectiveness factor can be calculated, and then the gaps in the financial performance can be estimated. The batch effectiveness factor considers complicated optimization procedures to generate ideal performance results, its equation is simple and easy to use. Generally, the batch effectiveness

factor can track the growth of a performance gap, as it can be measured at any time and for any operation in the process, provided there is an ideal performance target for that operation; therefore, this key performance measure can work online during the processing of the batch and give real-time results. If a gap starts to appear, an immediate improvement measure can be taken to alleviate the performance deviation; there is no such characteristic in other performance measures.

3.1.7. Step 7: Definition and Application of Improvement

Practices that are used to achieve ideal performance are needed, in order to define the required improvement actions. When poor performance is detected, the role of these practices is to define what actions should be taken to improve performance. Following that, the process engineers of manufacturing plants are encouraged to identify and analyze the requirement for improvement that exists among multiple performance alternatives, with the aim of closing the gap in understanding the effect of chemical engineering concepts in performance results. This model aims to close this gap by defining the process parameters by which the performance is influenced. Basic concepts are needed to understand the interaction between materials, equipment, and control parameters, such as temperature and speed of processing. Analytical techniques are also required to demonstrate an understanding of the ways to prevent failure to meet performance targets.

Below are some of the best manufacturing practices, utilized by world-class companies to meet increasingly demanding customer expectations and continually improve operating performance.

- *Lead Time Reduction*: world-class manufacturers are measuring and continually reducing manufacturing lead times. Time of non-value-adding steps in the manufacturing process is eliminated or minimized, to increase manufacturing velocity and therefore increase value to customer.
- *Process Reliability*: manufacturing improvement practices are in place to maximize equipment uptime and reduce variation in product quality. Effective methods and tools are also used for quick changeover, to improve equipment availability performance, and respond quickly and efficiently to changes in schedule and customer needs. Operating procedures and quality standards are consistent and manufacturing improvement practices are used to ensure sustainability.
- *Process Modeling and Simulation*: process modeling and simulation are used owing to two main needs [64, 65]: to gain better understanding of a specific process and to reduce the range of experimental work required for process development.
- *Performance Measurement*: key performance indicators and shop-floor goals are in place for each area, developed at the operating level and tied directly to plant goals.

Generally, operational performance excellence in each of the above areas will give the plant a real competitive edge.

3.2. Example of Practical Application

Two batch processes from a specialty chemicals plant are selected to evaluate the applicability of the developed operational effectiveness model. The actual field data of cycle time, yield, and batch quantity were monitored and recorded for manufacture of products A and B. Observations highlighted some bottleneck process steps in the process cycle time and also identified significant amounts of material waste during the handling and transfer of materials in these two processes. The bottlenecks in the cycle time steps were found mainly in the time needed to heat or cool the batch to the required temperature. Transfer of material has a subsequent longer time. The calculated results show that heat-transfer time, mixing, and material addition can be reduced significantly by the use of optimum operational conditions. Further, the time spent by samples in quality control can be eliminated by using online sampling techniques. The use of overhead tanks to reduce material addition time in the case-study plants has little support from the point of view of this research, as the cleaning process of the equipment is expensive and time-consuming. In addition, due to some exothermic reactions, some of the material requires a slow rate of addition, so the use of overhead

Table 2. Results of potential improvement in performance

	Product A		Product B	
Measure	Actual	Theoretical	Actual	Theoretical
Cycle time, h	12.24	6.39	9	5.2
Yield, %	92.2	94.7	97.36	100
Batch quantity, drums	17	varies	17	varies
Batch effectiveness factor, %	50.6	100	56.3	100

tanks will not decrease the time of addition in some steps of the process. The theoretical yield was also defined for the two manufacturing processes. The two values of the theoretical yield of both products showed that there are significant amounts of waste; this will affect the value of the actual yield. At the manufacturing plant, efforts are needed to eliminate the source of these material losses and increase the amount of product. Economic batch quantities at different assumed inventory consumption rates were calculated. The results showed that the financial gap will increase each time the difference between the actual and economic batch quantity increases. The effectiveness measures and the batch effectiveness factor were calculated; hence, financial analyses are performed, in order to evaluate performance gaps and the potential for improvement. As shown in Table 2, defining the actual performance for each effectiveness measure and defining the theoretical performance targets made it possible to calculate the batch effectiveness factor for the two processes. The financial gaps related to these values of the batch effectiveness factors are significant. These gaps in the operational and financial performance can still be closed, by eliminating the causes of performance gaps. The optimum performance was determined, but in this case was based on the available specifications of the system and the materials. These optimum results may improve further if a better system or raw materials were used.

The batch effectiveness factor proved to be an appropriate performance measure for the batch process at the chemical plant. Application to the selected case-study plant revealed significant losses in time, materials, and money in the existing actual processes; these losses can be eliminated by definition of the operational performance objectives and adoption of the operational effectiveness model, as defined in this article. The developed tools showed the improvement value, in terms of providing an understanding of how lower level performance improvements can have impact on a firm's ability to achieve its objectives. The improvement in the process time and yield of the final product caused an increase in the financial value of the product, thus creating financial savings. The improvement in the total cycle time of the process and the final product yield also supported excellent performance of the manufacturing plant. These improved results led to flexibility of operations, improved delivery performance, and less exposure to the hazards associated with the chemical processes. The achieved improvements will therefore result in the improvement of strategic KPIs, such as adherence to the production plan, OEE, OTIF, and value added per manufacturing employee. Consequently, the strategic KPIs at the top level of the manufacturing company will be supported and rise to the levels of performance excellence.

4. Conclusion

This article has introduced the requirement that a manufacturing organization needs to measure performance in chemical plants. Performance measurement frameworks, different categories of measures, target setting methods, and how to solve particular problems related to performance measurement are described. Defining the effective performance of operations in the chemical plants is vital, in order to achieve success in manufacturing businesses. To increase the effectiveness of operations within these plants, however, it is imperative that there is an understanding of the relationship between performance and the achievement of objectives. Development of new tools is also needed to assess the performance of the operation in chemical batch plants. In this context, the operational effectiveness model in this article was developed. It comprises a new

key performance indicator, called the "batch effectiveness factor", and the use of theoretical procedures to define performance targets. Defining the theoretical targets as optimum states of the process was found to be supportive to management, to define the routes to attain their objectives, investigate the root cause of any deviation, and understand the financial significance of being at the level of best-in-class operations.

5. References

1. M. Ahmad, R. Benson: *Benchmarking in the Process Industries*, IChemE, Rugby, UK. 1999.
2. P. Crosby: *Quality is Free*, McGraw Hill, New York 1978.
3. J. Juran, F. Gryna: *Quality Planning and Analysis*, McGraw Hill, New York 1993.
4. W. Skinner: "Manufacturing — Missing Link in Corporate Strategy", *Harvard Business Rev*. **50** (1969) no. 3, 136 – 145.
5. A. Neely, H. Richards, J. Mills, K. Platts, M. Bourne: "Designing Performance Measures: A Structured Approach", *Int. J. Operation Production Management*, **17/11** (1997), 1131 – 1152.
6. J. Lingle, W. Schiemann: "From Balanced Scorecard to Strategic Gauges: Is Measurement Worth It?", *Manag. Rev.*, March (1996), 56 – 61.
7. J. Harbour: *The Basics of Performance Measurement*, Quality Resources, New York, 1997.
8. N. Oliver, R. Delbridge, D. Jones, J. Lowe: "World Class Manufacturing: Further Evidence in the Lean Production Debate", *Br. J. Manag.* **5** (1994) Special Issue, June, S53 – S63.
9. V. Gilgeous, M. Gilgeous: "A Framework for Manufacturing Excellence", *Integrated Manufact. Sys.* **10** (1999) no. 1, 33–44.
10. R. Schonberger: *World Class Manufacturing: The Lessons of Simplicity Applied*, The Free Press, New York, NY, 1986.
11. M. Ahmad, N. Dhafr: "Measuring Manufacturing Performance of Process Plants", Proceedings of the 13th International Conference on Flexible Automation Intelligent Manufacturing, FAIM 2003, University of South Florida, USA, 2003, pp. 570 – 578.
12. M. Zairi: *Measuring Performance for Business Results*, Chapman & Hall, London 1994.
13. J. Wisner, S. Fawcett: "Linking Firm Strategy to Operating Decisions Through Performance Measurement", *Prod. Inventory Manag. J.* **32** (1991) no. 3, 5 – 11.
14. P. Swamidass, W. Newell: "Manufacturing Strategy, Environmental Uncertainty, and Performance: A Path Analytical Model", *Manag. Sci.* **33** (1987) no. 3, pp. 509–534.
15. *Webster's American English Dictionary*, Collins, London, 1990.
16. S. Wheelwright: "Reflecting Corporate Strategy in Manufacturing Decisions", *Business Horizons*, (1978) February, 57 – 66.
17. M. Ahmad, N. Dhafr, R. Benson: "Theoretical Targets for Performance of Manufacturing Shop Floor Operations", Proceedings of the 14th International Conference on Flexible Automation and Intelligent Manufacturing, FAIM2004, Toronto, Canada, 2004.
18. M. Ahmad, N. Dhafr, R. Benson, B. Burgess: "Model for Establishing Theoretical Targets at the Shop Floor Level in Specialty Chemicals Manufacturing Organizations", *Robotics Comput.-Integrated Manuf. J.* **21/4** (2005), 391 – 400.
19. B. Ganapathy, C. Goh: "A Hierarchical System of Performance Measures for Concurrent Engineering", *Concurr. Eng. Res. Appl.* **5** (1997) no. 2, 137 – 143.
20. L. Lynch, K. Cross: *Measure Up – the Essential Guide to Measuring Business Performance*, Mandarin, London, 1991.
21. R. Kaplan: "Yesterday's Accounting Undermines Production", *Harvard Business Rev.* (1984) July–August, 95 – 101.
22. P. Walsh: "Finding Key Performance Drivers: Some New Tools", *Total Qual. Manag.* **7** (1996) no. 5, 509 – 520.
23. M. Epstein, J. Manzoni: "Implementing Corporate Strategy: From Tableau de Bord to Balanced Scorecard", *Eur. Manag. J.* (1998) April, 190 – 203.
24. R. Kaplan, D. Norton: "The Balanced Scorecard: Measures that Drive Performance", *Harvard Business Rev.* (1992) Jan – Feb, 71 – 79.
25. A. Neely, C. Adams, M. Kennerly: *The Performance Prism: the Scorecard for Measuring and Managing Business Success*, Financial Times-Prentice Hall, London 2002.
26. D. Keegan, R. Eiler, C. Jones: "Are your Performance Measures Obsolete?" *Manag. Account. (US)* **70/12** (1989) June, 45 – 50.

27. N. Slack, S. Chambers, R. Johnston: *Operations Management, Prentice Hall*: Financial Times, London 2004.
28. G. Najarian: "Performance Measurement: Measure the Right Things", *J. Manuf. Sys.* (1993) Sept., 54 – 57.
29. B. Maskell: *Performance Measurement for World Class Manufacturing: A Model for American Companies*, Productivity Press, Inc. Cambridge, MA, 1991, pp. 305 – 329.
30. J. Dixon, A. Nanni, T. Vollmann: *The New Performance Challenge: Measuring Operations for World-Class Competition*, Irwin, Homewood, IL, 1990.
31. A. Ghalayini, J. Noble: "The Changing Basis of Performance Measurement", *Int J. Operat. Prod. Manag.* **16** (1996) no. 8, pp. 63 – 80.
32. L. Tatikonda: "We Need Dynamic Performance Measures", *Manag. Account. (US)* (1998) Sept., 49 – 53.
33. M. Ahmad, N. Dhafr: "Establishing and Improving Manufacturing Performance Measures", Proceedings of the 11th International Conference on Flexible Automation Intelligent Manufacturing. FAIM, Dublin City University, Ireland, 2001.
34. M. Ahmad, N. Dhafr: "Establishing and Improving Manufacturing Performance Measures" *Robotics Comput.-Integr. Manuf. J.* **18/3** (2002) June, 171 – 176.
35. S. Nakajima: *Introduction to TPM — Total Productive Maintenance*, Productivity Press, Cambridge, MA, USA, 1988.
36. L. Fortuin: "Performance Indicators – Why, Where, and How?", *Eur. J. Operat. Res.* **34** (1988) no. 1, 1 – 9.
37. P. Walsh: "Targets and How to Assess Performance Against Them", *Benchmarking Int. J.* **7** (2000) no. 3, 185.
38. G. Medori: Establishment of Non-Financial Performance Measures in Medium to Large–Sized Organizations Consistent with World-Class Manufacturing Objectives, PhD Thesis, Coventry University, 1998.
39. G. White: "A Survey and Taxonomy of Strategy-Related Performance Measures for Manufacturing", *Int J. Operat. Prod. Manag.* **16** (1996) no. 3, 42 – 61.
40. I. Mujtaba: *Batch Distillation: Design and Operation*, Imperial College Press, London 2004, p. 33.
41. U. Diwekar: *Batch Distillation. Simulation, Optimal Design and Control*, Taylor Francis, Washington, DC, 1994.
42. A. Steinbach, R. Winkenbach: "Choose Processes for their Productivity", *Chem. Eng.* (2000) April, 94 – 105.
43. C. Waters: *Inventory Control and Management*, Wiley, New York 1992.
44. J. Womack, D. Jones: *Lean Thinking: Banish Waste and Create Wealth in your Corporation*, Simon Schuster, New York 1996.
45. J. McEwan: "How to Predict Batch-Reactor Heating and Cooling", *Chem. Eng.* (1989) May,179.
46. I. Tosun, I. Aksahin: "Predict Heating and Cooling Times Accurately", *Int J. Operat. Prod. Manag.* (1993) Nov., 183.
47. V. Mahajani: "Estimate Batch Reactor Heating or Cooling Time", *Chem. Eng.* (1996) Sept., 143.
48. D. Kern: *Process Heat Transfer*, McGraw Hill, New York 1950.
49. W. McCabe, J. Smith, P. Harriott: *Unit Operations of Chemical Engineering*, 5th Ed, McGraw-Hill, New York 1993, p. 245.
50. R. Perry, D. Green, *Perry's Chemical Engineers' Handbook*, 7th Ed, McGraw Hill, New York 1998.
51. J. Fourier: *The Analytical Theory of Heat*, Dover Publications, New York, 1955.
52. K. Whitten, K. Gailey, R. Davis: *General Chemistry with Qualitative Analysis*, 4th ed., Saunders College Publishing, New York 1992, p. 616.
53. S. Radel, R. Navidi, *Chemistry*, 2nd Edition, West Publishing Company, St. Paul, Minneapolis 1994, p. 632.
54. Y. Vemuri: Real-Time Optimization of Semi-Batch Reactors, PhD Thesis, the Florida State University, College of Engineering, 2004.
55. H. Fogler: *Elements of Chemical Reaction Engineering*, Prentice Hall, New York 1999, pp. 290 – 291.
56. Denbigh, K. J. Turner: *Chemical Reactor Theory*, Cambridge University Press, 1971, pp. 103 – 104.
57. J. Smith: *Chemical Engineering Kinetics*, McGraw-Hill, New York 1981.
58. D. Himmelblau: *Basic Principles and Calculations in Chemical Engineering*, Prentice-Hall International, New York 1992, p. 71.
59. J. Evans: *Applied Production and Operations Management*, 4th ed., West Publishing Company, St. Paul, Minneapolis 1992, p. 429.
60. R. Wild: *Production and Operations Management*, 5th ed., Continuum, London 1995, p. 409.

61. J. Dilworth: *Operations Management: Providing Value in Goods and Services*, 3rd ed., South-Western College Publishers, Cincinnati, Ohio 1999, p. 423.
62. F. Harris: *Operations and Cost*, A. W. Shaw Co. Factory Management Series, Chicago, 1915.
63. M. Ahmad, N. Dhafr: "Setting Theoretical Targets for Improvement of Manufacturing in Speciality Chemicals Industry", Proceedings of the 7th World Congress of Chemical Engineering, Glasgow, 2005.
64. "Why Process Modelling?", *Chem. Eng.* (1987) Nov., 9.
65. "The Benefits of Simulation?", *Chem. Eng.* (2003) July, 1.

Epothilones

BERND BUCHMANN, Medicinal Chemistry, Schering AG, Berlin, Germany
ULRICH KLAR, Medicinal Chemistry, Schering AG, Berlin, Germany

1. Introduction and History 1
2. Physical Properties 1
3. Syntheses 2
3.1. Synthesis of Epothilones A and C According to SCHINZER 3
3.2. Synthesis of Epothilones A and C According to DANISHEFSKY 5
3.3. Synthesis of Epothilones B and D According to NICOLAOU 5
3.4. Synthesis of Epothilone D According to DANISHEFSKY 6
4. Occurrence and Biosynthesis 6
5. Pharmacological Effects and Uses . . 7
6. References 13

Epothilones represent a novel class of anticancer drugs which inhibit the cell cycle by interfering with microtubules. Their chemical and pharmacological properties as well as several synthetic approaches to some of the natural compounds are described.First results from clinical trials with natural epothilones as well as novel derivatives are promising indicating that epothilones have the potential to become a cornerstone in future cancer therapy.

1. Introduction and History

In 1991, the first two members of a new class of natural products (epothilones A and B) were described in a patent application by the groups of HÖFLE and REICHENBACH [17]. Both compounds were isolated from the myxobacterial strain *Sorangium cellulosum* originated from a soil sample collected in the bed of the Zambesi river in South Africa. Beside several potential uses as plant protecting or antifungal agents, also their cytotoxic or immunosuppressive properties were claimed.

In 1995, BOLLAG et al. at Merck Research Laboratories reported that epothilones exert their high antiproliferative activity by the stabilization of microtubules. Up to this time, taxanes had been the only class of compounds addressing this type of mechanism (see → Cancer Chemotherapy, Chap. 7). However, in contrast to taxanes, epothilones possess the ability to maintain their cytotoxic activity also in tumor cells that have developed marked resistancies against widely used therapeutic agents like Taxol® or anthracyclines. A better water solubility of epothilones compared with that of taxanes offers higher flexibility for the development of formulations that avoid the use of cremophore to which is attributed the amplification of side effects of Taxol®. Beside these potentially beneficial aspects, the most potent natural compound epothilone B still displays significant toxicities at therapeutically relevant doses that result in a narrow therapeutic window. Therefore, an epothilone analog possessing an improved therapeutic window would be highly desirable for clinical applications. The development of such a compound has stimulated intensive research activities in chemistry, pharmacology, and medicine [18, 19].

2. Physical Properties

Epothilones A to F are solids. They are well characterized by different methods such as ^{1}H and ^{13}C NMR, UV, IR, and mass spectrometry. Some physical properties are given in Table 1. Epothilones B and D are commercially available.

Stability. The pure compounds are quite stable but it is recommended to store the compounds at lower temperature (−20 °C to 4 °C). In solution, the stability depends on the pH. Under basic conditions, the lactone ring is opened,

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.k09_k01.pub2

Table 1. Physical properties of epothilones A to F

Epothilone	Formula	M_r	Appearance	*mp*, °C	$[\alpha]_D^{22}$ (c in methanol)	λ_{max} (methanol), nm	ε, L mol^{-1}cm^{-1}
A	$C_{26}H_{39}NO_6S$	493.66	colorless crystals	95 (ethyl acetate–hexane)	−47.1 ° (1.0)	211, 249	17800, 12500
B	$C_{27}H_{41}NO_6S$	507,68	colorless crystals	93–94 (ethyl acetate)	−35.0 ° (0.7)	211, 249	18600, 14100
C	$C_{26}H_{39}NO_5S$	477.66	colorless amorphous solid	–	−74.4 ° (2.0)	213, 248	16200, 12500
D	$C_{27}H_{41}NO_5S$	491.68	colorless crystals	120–121 (ethanol–water)	−61.3 ° (2.5)	210, 248	18400, 13200
E	$C_{26}H_{39}NO_7S$	509.66	colorless amorphous solid	–	−37.1 ° (0.56)	212, 247	17800, 13000
F	$C_{27}H_{41}NO_7S$	523.68	colorless amorphous solid	–	−27.4 ° (0.46)	212, 249	19100, 13800

in acidic solutions the epoxide can be cleaved by hydrolysis.

The metabolic stability of epothilone B has been intensively studied. Lactone and epoxide hydrolysis is the most prominent metabolic pathway [20].

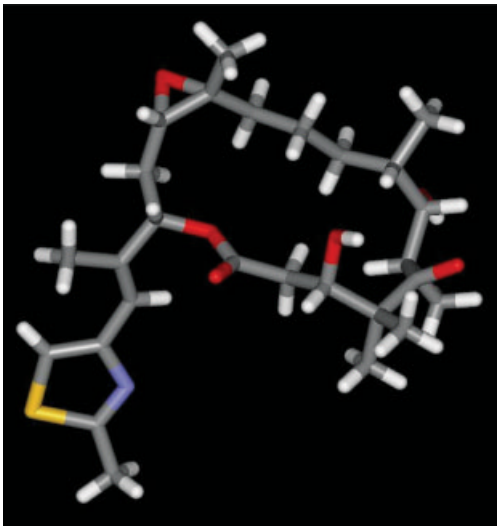

Figure 1. X-ray structure of epothilone B

Epothilones are 16-membered macrocyclic lactones containing two hydroxyl groups, one carbonyl group, and a *Z*-configurated double bond which can also be oxidized to an epoxide. Attached to this ring system is an unsaturated side chain bearing a thiazole. From these molecular features, the name EPOTHILONE was created: EPOxide, THIazole, LactONE. The compounds possess 7 to 8 stereogenic centers and are isolated in optically pure form. The relative and absolute configuration of epothilones A and B (Fig. 1) as well as their three-dimensional conformation were undoubtedly confirmed by X-ray crystallography [21, 22].

The chemical structures of epothilones A to F are shown in Table 2. Beside the IUPAC nomenclature, an alternative numbering became generally accepted which is depicted for epothilone D as an example (Fig. 2). This counting is also applied for the epoxide containing epothilones and will be used in the following description of the syntheses too.

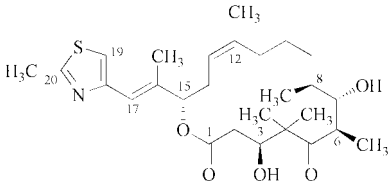

Figure 2. Epothilone D with alternative numbering

3. Syntheses

A large number of total syntheses have been described in numerous review articles and monographs [2 – 5, 12, 15, 16]. Common to each approach is a modular and highly convergent synthesis based on different building blocks. Beside the natural compounds, these strategies also allow the preparation of various analogs. The methodologies developed by the different research groups have found entry into the industrial development.

In Table 3, four different strategies to construct the 16-membered macrocycle are outlined. Exemplarily, a synthesis will be described for each strategy.

Table 2. Chemical structures of epothilones A to F, their CAS numbers, and IUPAC names

Structure	Name, CAS No.	IUPAC name
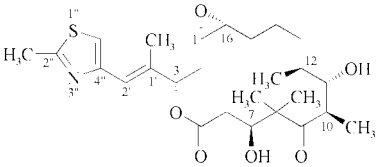	epothilone A[*152044-53-6*]	(1*S*,3*S*(*E*),7*S*,10*R*,11*S*,12*S*,16*R*)-4,17-dioxabicyclo[14.1.0]heptadecane-5,9-dione, 7,11-dihydroxy-8,8,10,12-tetramethyl-3-[1-methyl-2-(2-methyl-4-thiazol-yl)ethenyl]
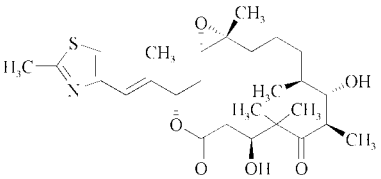	epothilone B [*152044-54-7*]	(1*S*,3*S*(*E*),7*S*,10*R*,11*S*,12*S*,16*R*)-4,17-dioxabicyclo[14.1.0]heptadecane-5,9-dione, 7,11-dihydroxy-8,8,10,12,16-penta-methyl-3-[1-methyl-2-(2-methyl-4-thiazolyl)ethenyl]
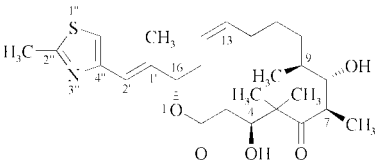	epothilone C [*186692-73-9*]	(4*S*,7*R*,8*S*,9*S*,13*Z*,16*S*(*E*))-oxacyclo-hexadec-13-ene-2,6-dione, 4,8-dihydroxy-5,5,7,9-tetramethyl-16-[1-methyl-2-(2-methyl-4-thiazol-yl)ethenyl]
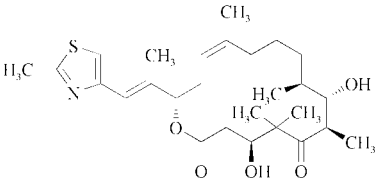	epothilone D [*189453-10-9*]	(4*S*,7*R*,8*S*,9*S*,13*Z*,16*S*(*E*))-oxacyclo-hexadec-13-ene-2,6-dione, 4,8-dihydroxy-5,5,7,9,13-pentamethyl-16-[1-methyl-2-(2-methyl-4-thiazol-yl)ethenyl]
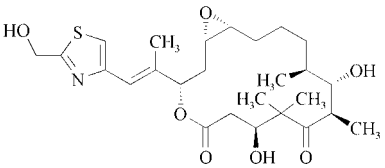	epothilone E [*201049-37-8*]	(1*S*,3*S*(*E*),7*S*,10*R*,11*S*,12*S*,16*R*)-4,17-dioxabicyclo[14.1.0]heptadecane-5,9-dione, 7,11-dihydroxy-3-[2-[2-(hydroxymeth-yl)-4-thiazolyl]-1-methylethenyl]-8,8,10,12-tetramethyl
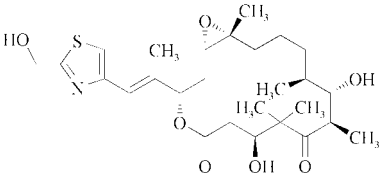	epothilone F [*208518-52-9*]	(1*S*,3*S*(*E*),7*S*,10*R*,11*S*,12*S*,16*R*)-4,17-dioxabicyclo[14.1.0]heptadecane-5,9-dione, 7,11-dihydroxy-3-[2-[2-(hydroxymeth-yl)-4-thiazolyl]-1-methylethenyl]-8,8,10,12,16-pentamethyl

3.1. Synthesis of Epothilones A and C According to SCHINZER

One of the pioneers who successfully synthesized epothilone A was the group of SCHINZER et al. [23, 24].

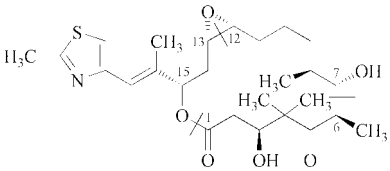

Figure 3. Structure of epothilone A with the strategic bond formations

Table 3. Overview on different strategies to construct the macrocycle

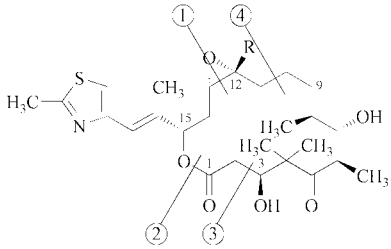

R = H: Epothilone A
R = CH_3: Epothilone B

Total syntheses of epothilones	Group	Cyclization strategy
A	Nicolaou	olefine metathesis 1
A	Schinzer	olefine metathesis 1
A	Nicolaou	solid-phase olefine metathesis 1
A + B	Nicolaou	macrolactonization 2
A + B	Danishefsky	macrolactonization 2
A + B	Danishefsky	aldol condensation 3
B	Danishefsky	olefine metathesis 1 + 4

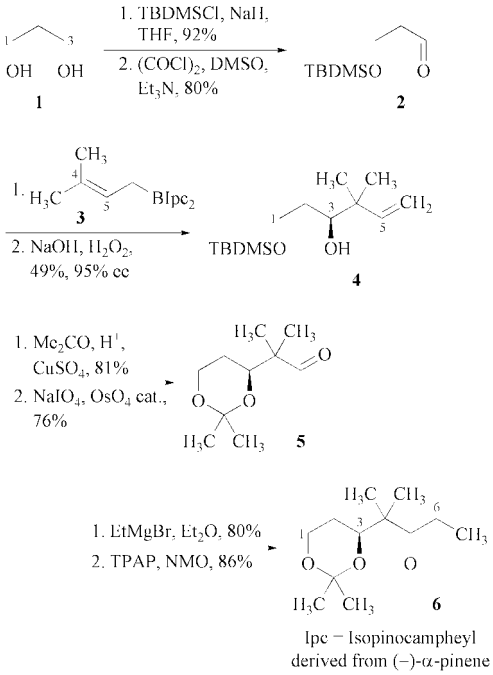

Figure 4. Synthesis of fragment C1-C6

Synthesis of fragment C1-C6 starts from commercially available propane-1,3-diol in which one hydroxyl group is selectively protected (Fig. 4). After the remaining hydroxyl group has been oxidized to aldehyde **2**, a chain elongation with the optically pure borane **3** yields the chiral alcohol **4**. After formation of the acetonide, the double bond is cleaved to give aldehyde **5**, which is transformed via a Grignard reaction followed by an oxidation to the desired fragment **6**.

Synthesis of fragment C7-C12 is achieved by a diastereoselective alkylation of **9** followed by reductive cleavage and subsequent oxidation (Fig. 5). Compound **9** is obtained from reaction of the chiral auxiliary **7** with the acid chloride **8**.

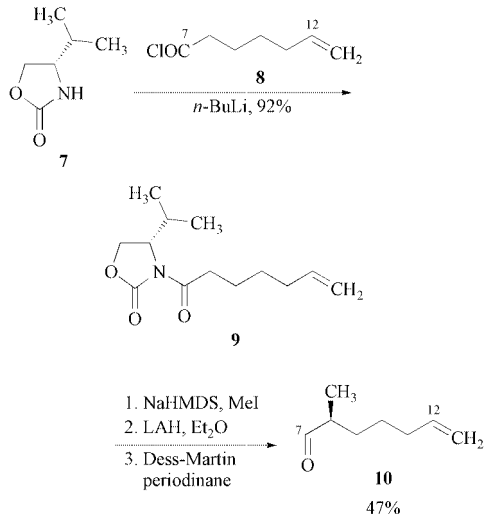

Figure 5. Synthesis of fragment C7-C12

Synthesis of fragment C13-C15 starts from the known aldehyde **2** already used in the synthesis of fragment C1-C6 (Fig. 6). Grignard reaction with **11** generates the racemic alcohol **12** from which the optically pure compound **13** is obtained via a kinetic resolution procedure according to Sharpless. Silylation and oxidative cleavage of the double bond yields ketone **14** which is subjected to a Wittig-Horner type reaction with phosphonate **15**. Selective deprotection of the primary silyl ether gives alcohol **16** that is oxidized to the aldehyde and subjected to a further Wittig reaction to form the desired fragment **17** after desilylation.

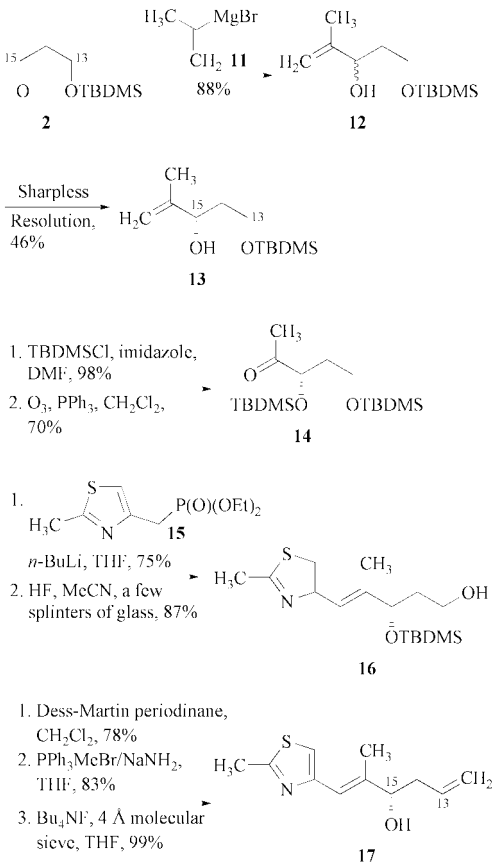

Figure 6. Synthesis of fragment C13-C15

Now, compounds **6** and **10** are connected in an aldol reaction by which two additional stereocenters are generated with high stereoselectivity (Fig. 7). After the acetonide is removed, triol **18** is persilylated. The most labile primary silyl ether is selectively cleaved and the resulting alcohol oxidized to the carboxylic acid **19**.

To complete the carbon skeleton, acid **19** is esterified with alcohol **16** to give the ester **20**. The formation of the 16-membered ring is accomplished using the ring-closing metathesis (RCM) as key reaction yielding a 1:1 mixture of the *E*/*Z* double bond isomers in a total yield of 94%. After separation, the protecting groups in **21** are removed to give epothilone C. Stereoselective epoxidation of the double bond furnished epothilone A along with minor amounts of its diastereomeric epoxide.

3.2. Synthesis of Epothilones A and C According to Danishefsky

In contrast to the afore described synthesis of epothilone A, Danishefsky et al. have chosen a quite different approach to construct the macrocyclic ring system (Fig. 8) [25, 26].

Synthesis of fragment C3-C11 (**23**) is accomplished in a multistep sequence starting from the known optically active aldehyde **22** while fragment C12-C2 (**25**) is synthesized in several steps starting from the commercially available (R)-glycidol **24**. The Suzuki-type coupling of the two fragments is achieved by the borane formed from **23** with the iodo olefine **25**.

As key intermediate for ring closure the enolate of acetate **26** is used in a stereoselective aldol reaction with the aldehyde. The cyclization proceeds smoothly since an enolization of the aldehyde is prevented by the adjacent quaternary center.

After deprotection a selective silylation of the 3- and 7-hydroxyl groups enables the oxidation of the 5-hydroxyl group to the ketone. The final deprotection leads to epothilone C which can be transformed into epothilone A as already described.

3.3. Synthesis of Epothilones B and D According to Nicolaou

In this first synthesis of epothilones B and D, a macrolactonization of **31** has been chosen as key step for the ring construction (Fig. 9) [27].

Fragment C7-C15 is produced via an unselective Wittig reaction of **28** with ketone **29**

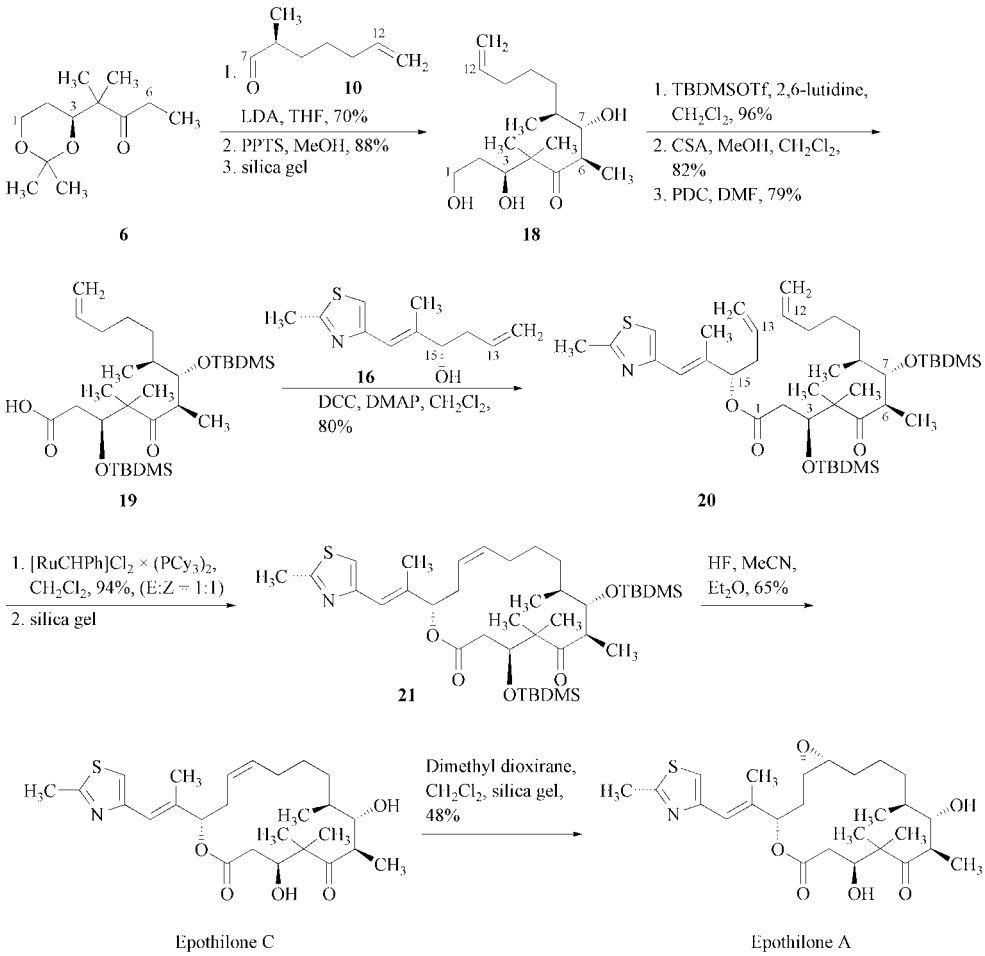

Figure 7. Final synthesis of epothilones A and C

followed by selective desilylation and oxidation to the aldehyde **30**. Similar to the Schinzer approach, the C1-C6 fragment is introduced via a stereoselective aldol reaction. Macrolactonization under Yamaguchi conditions, chromatographic removal of the 12*E*-isomer and subsequent cleavage of the remaining protecting groups gives epothilone D, which is further oxidized to epothilone B.

3.4. Synthesis of Epothilone D According to Danishefsky

Another RCM approach to connect center C10 with C11 has been developed for the synthesis of epothilone D by Danishefsky et al. (Fig. 10) [28].

The required key compound **34** is prepared by a stereoselective aldol reaction between acetate **32** and ketoaldehyde **33**. After macrocyclization using the Grubbs-II catalyst the synthesis to epothilone D is accomplished by a reductive cleavage of the protecting group followed by a regioselective imine reduction of the less substituted double bond.

4. Occurrence and Biosynthesis

From the myxobacterial strain *Sorangium cellulosum*, epothilones are produced with yields of about 20 mg L^{-1}. The 16 hour doubling time makes a production in this organism economically less attractive. Therefore, this gene cluster

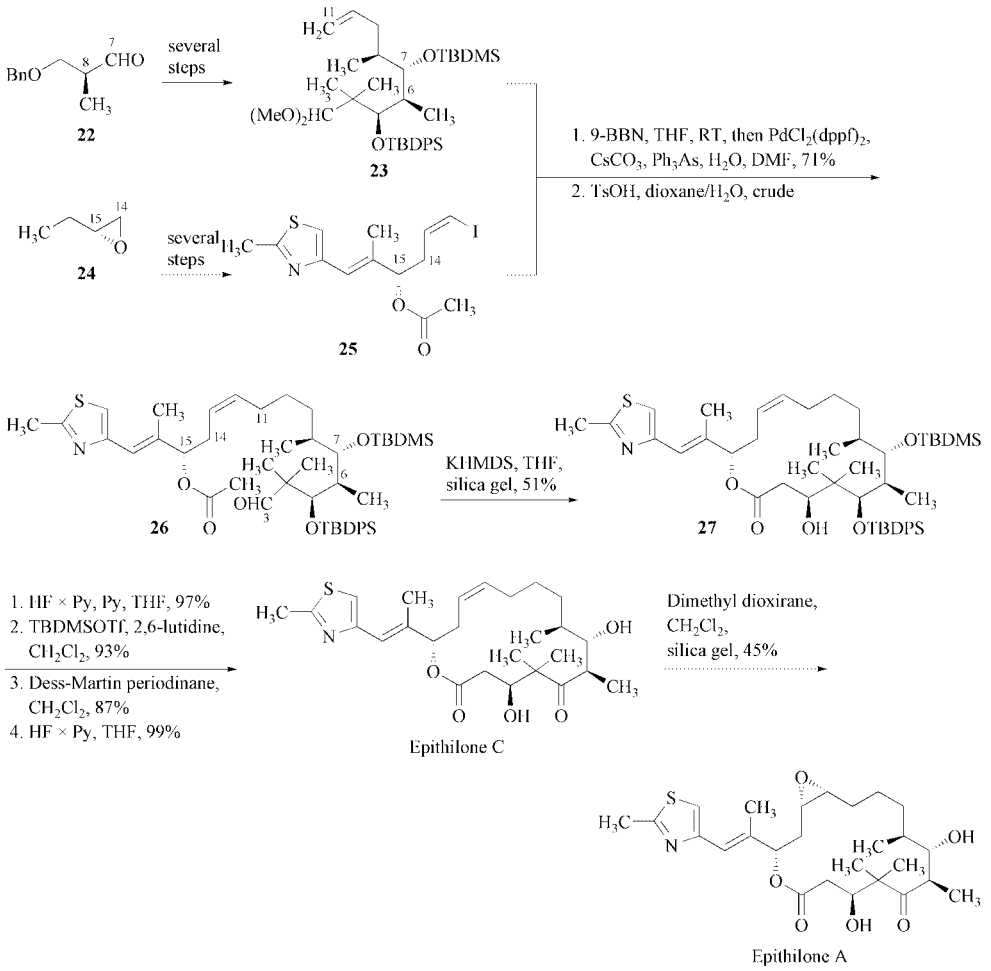

Figure 8. Synthesis of epothilones A and C according to DANISHEFSKY et al.

has recently been cloned and expressed in *Myxococcus xanthus*, a related host more amenable to genetic manipulation with a significantly shorter doubling time.

The biosynthetic pathway for the production of the epothilones in *S. cellulosum* has been elucidated and studied in detail [29 – 32]. The biosynthetic gene cluster (epo) has been isolated, sequenced, and found to code for six multifunctional proteins (Fig. 11). These proteins comprise nine polyketide (PKS) synthases (epoA, epoC - epoF) and one nonribosomal peptide (NRPS) synthetase (epoB). Finally, a P450 epoxidase converts epothilones C or D into the epoxide-bearing epothilones A or B, respectively.

In the meantime, mutant strains lacking the P450 epoxidase have been created to produce predominantly epothilones C and D in a 1:4 ratio. Further optimizations addressing culture medium, fermentation conditions and purification procedures led to a multigram synthesis of epothilone D [33 – 35].

5. Pharmacological Effects and Uses

The mitotic spindles that control the movement of chromosomes during cell division are composed of microtubules. Microtubules are formed from helical arrays of alternating α and β subunits of the protein tubulin. The formation of microtubules involves the assembly of tubulin

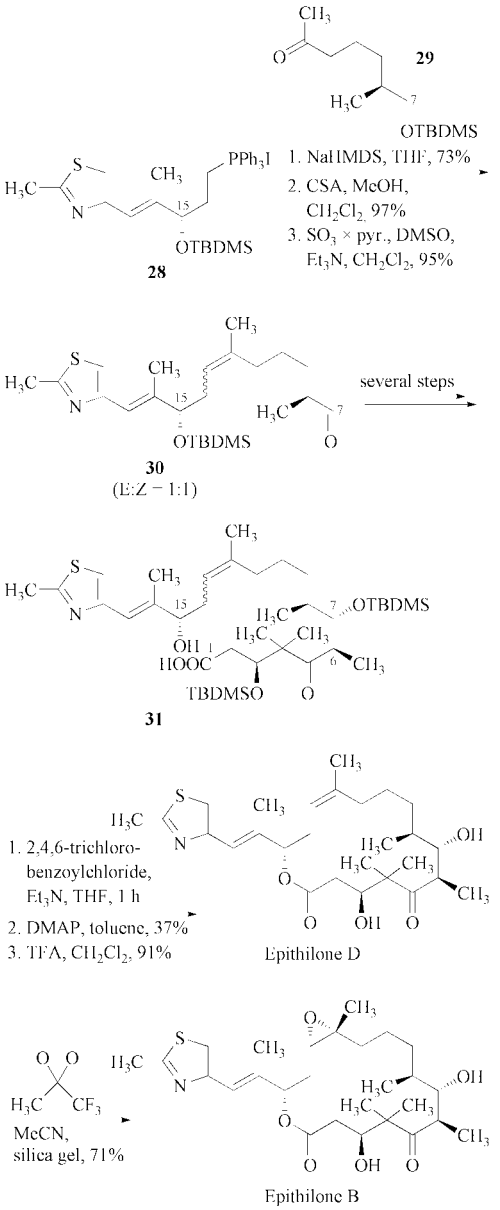

Figure 9. Synthesis of epothilones B and D according to NICOLAOU et al.

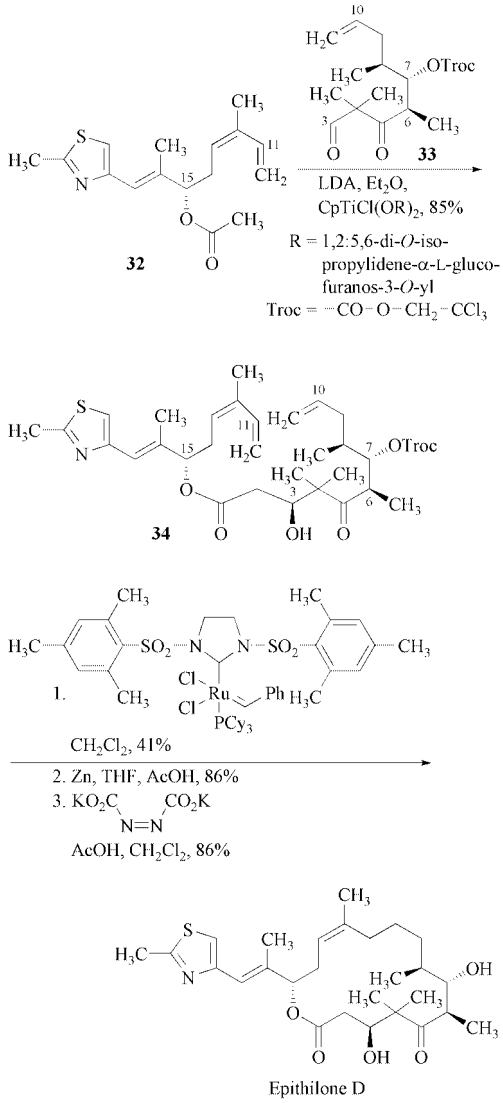

Figure 10. Synthesis of epothilones D according to DANISHEFSKY et al.

subunits bound to guanosine triphosphate, providing the energy for microtubule movement. Microtubules play a central role in cell structure, signaling, movement, and transport, as well as mitosis and meiosis. Disrupting these functions, therefore, leads to mitotic arrest followed by apoptosis.

Microtubules are recognized as clinically relevant targets for tumor therapy based on their integral role in cell division. Several naturally occurring cytotoxic agents disrupt cell division by affecting microtubule formation in a variety of ways. Vinca alkaloids prevent tubulin polymerization, while taxanes promote polymerization and disrupt microtubule dynamics. One of

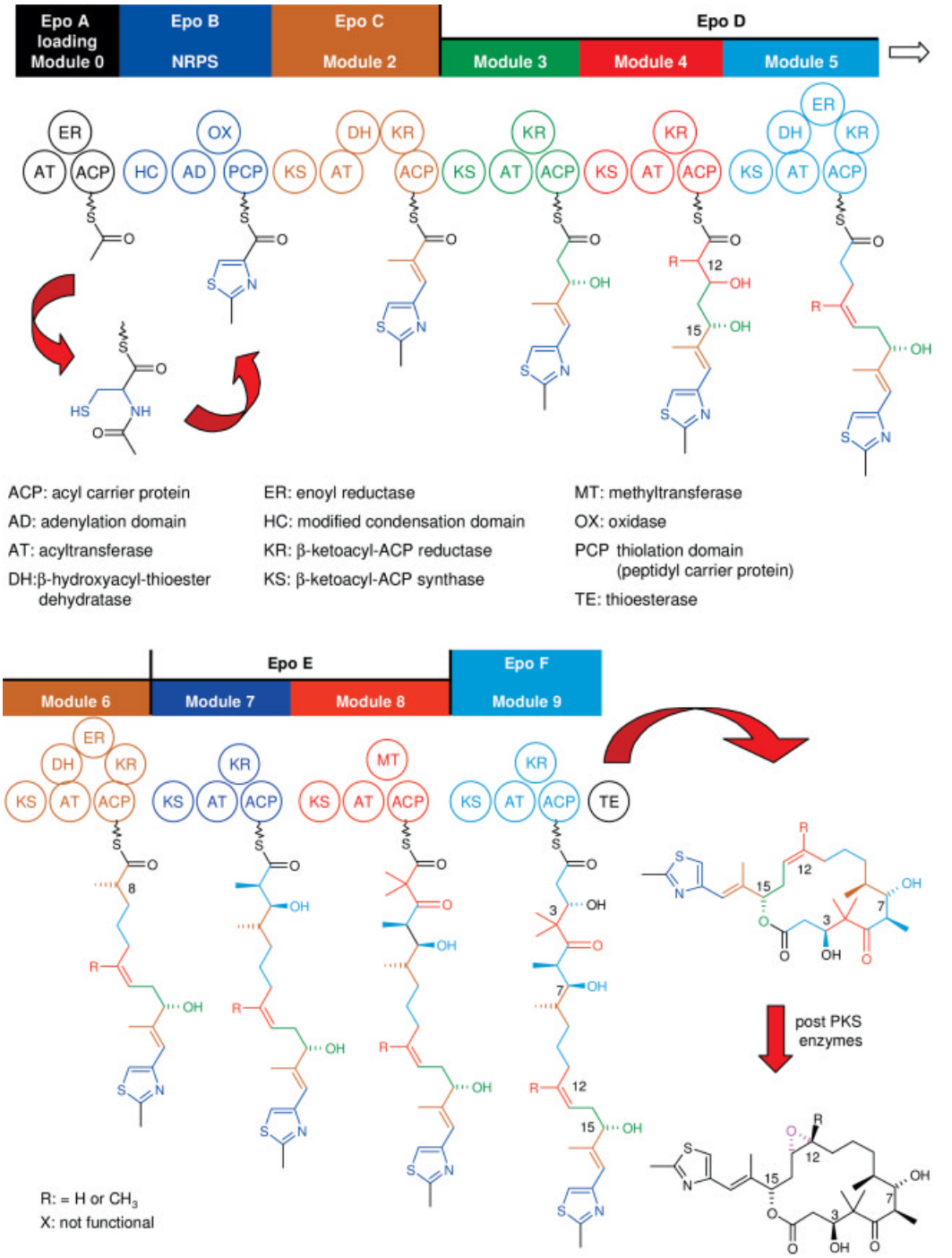

Figure 11. Biosynthetic pathway for the production of epothilones in *Sorangium cellulosum*

Table 4. Epothilone B inhibits the growth of a wide range of human carcinoma and drug-resistant human cancer cell lines [44]

		IC_{50} for net growth inhibition, nM		
Human carcinoma cell lines		**Epothilone A**	**Epothilone B**	**Paclitaxel**
Lung	A549	2.67	0.23	3.19
	NCI-H460	3.17	0.29	5.65
Colon	HCT116	2.51	0.32	2.79
	SW620	2.0	0.1	0.2
Breast	MDA-MB-231	1.31	0.13	1.43
	BT-20	3.63	0.13	1.83
	ZR-75-1	3.80	0.64	3.60
	MCF-7	1.49	0.18	1.80
Prostate	DU145	4.86	0.31	4.14
	PC-3M	4.27	0.52	4.77
Bladder	T-24	0.25		4.40
Epidermoid	A-431	2.35	0.26	1.66
	KB-31	2.1	0.19	2.31
Leukemia	CCRF-CEM	2.7	0.35	2.1
Ovarian	1A9	2.0	0.06	2.0
Drug-resistant human cancer cell lines				
Colon	SW620AD-300	3.0	0.3	250
	HCT-15	3.6	0.41	136
Breast	MCF-7/ADR	27.5	2.92	9105
Epidermoid	KB-8511	1.9	0.19	533
Ovarian	1A9PTX22	3.0	0.1	43
T cell ALL	CCRF-CEM/VBL100	20.0	2.1	4140

the limitations of most taxanes and many other cytotoxic agents, however, is the development of resistance.

Approximately 75 % of cancers are either intrinsically unresponsive or develop resistance to treatment [36]. However, there is a new class of agents, the epothilones, which have a similar mechanism of action as Taxol® but exert an antiproliferative effect on cells that exhibit multidrug resistance. Unlike traditional drugs, such as colchicines, vincristine, or vinblastine, which stabilize tubulin but destabilize microtubules, taxanes and epothilones stabilize preformed microtubules. Both mechanisms have the potential to cause cell death; however, only epothilones possess the potential to overcome multidrug resistance in a variety of tumors [37, 38].

Similar to other tubulin-targeting small molecules, epothilones bind to the β subunit of heterodimeric tubulin, causing enhanced polymerization followed by hyperstabilization of the formed microtubules. Electron crystallographic structures of tubulin with epothilone A [39] or paclitaxel [40] have identified a common binding pocket for both compounds. Since the dynamic formation and dismantling of microtubules is necessary for cell division, epothilones induce G2/M mitotic arrest and apoptosis. Epothilones have shown broad potential as antitumor agents based on this unique mechanism of action [9].

Epothilones maintain potency against tumor cells with acquired resistance. Drug resistance is often associated with overexpression of the P-glycoprotein efflux pump, which actively removes drugs from affected cells. This process can be mimicked in the laboratory by exposing cells to increasing drug concentrations and selecting those that survive. Using this model, cells that are 45-times more resistant to paclitaxel than the parent cell line can be obtained in just 5 months [41]. Most epothilones possess the potential to exert an antiproliferative effect on cells that exhibit multidrug resistance, maintaining potency after in vitro selection processes that induce paclitaxel resistance.

Among the most prominent naturally occurring epothilones (types A-D), epothilone B shows the highest activity in vitro and in vivo. Epothilone B inhibits the growth of a large number of human cancer cell lines in vitro, representing most solid tumor types (Table 4). Growth inhibition also closely correlates with the induction of apoptosis, as indicated by the integrity of the plasma membrane. It appears that epothilone B is retained inside cells, leading to high intracellular concentrations that facilitate effective inhibition of tubulin polymerization.

Table 5. Epothilone compounds currently under clinical investigation

Structure	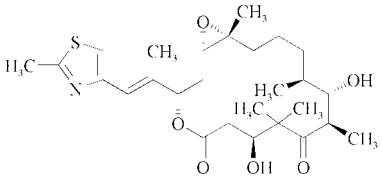	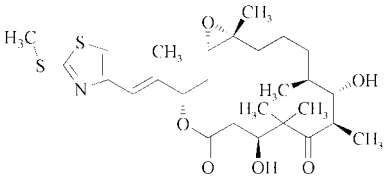
Name(s)	epothilone B, patupilone	
Code	EPO 906	ABJ 879
Generation	first	second
Company	Novartis	Novartis
Status of development in 2005	phase II	phase I
Cancer types	colon, ovarian, renal	
Stability	substrate for esterases (lactone)	substrate for esterases (lactone)
Pharmacokinetics in humans	$T_{1/2}$= 60h	
Formulation	PEG 300	
Synthesis	recombinant	partial synthesis from epothilone B
Structure	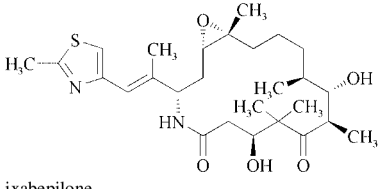	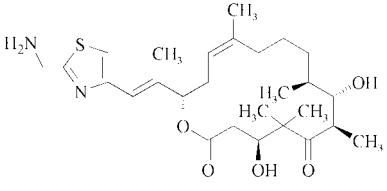
Name(s)	ixabepilone	
Code	BMS 247550	BMS 310705
Generation	second	second
Company	Bristol-Meyrs Squibb	Bristol-Meyrs Squibb
Status of development in 2005	phase III	phase I (since 2004 no further development reported)
Cancer types	breast, NSCLC, colon, prostate, gastric, ovarian	
Stability	deactivation by intramolecular triggered epoxide opening [45]	substrate for esterases (lactone)
Pharmacokinetics in humans	$T_{1/2}$= 36h	$T_{1/2}$= 33h
Formulation	cremophore EL; lyophilisate	not required (water soluble)
Synthesis	partial synthesis from epothilone B	partial synthesis from epothilone B

Table 5. (Continued)

Structure			
Name(s)	epothilone D, desoxy-epothilone B		
Code	KOS 862	KOS 1584	ZK-EPO
Generation	first	second	third
Company	Kosan, Roche	Kosan, Roche	Schering AG
Status of development in 2005	phase II	phase I	phase II
Cancer types	breast, NSCLC, colon		various solid tumors
Stability	substrate for esterases (lactone)	substrate for esterases (lactone)	substrate for esterases (lactone)
Pharmacokinetics in humans	$T_{1/2}$ = 5–10h		$T_{1/2}$ = 48–96h
Formulation	cremophore EL	HP-ß-CD-clathrate lyophilisate	no cremophore
Synthesis	recombinant	recombinant	total synthesis

Despite a promising profile in vitro, epothilone B is highly toxic in animal models in which different human cancer cell lines were xenografted to rodents. Furthermore, in human tumor cell lines that overexpress P glycoprotein, the addition of verapamil to block the efflux pump enhances the activity of epothilone B about 10-fold, indicating that epothilone B is still a substrate for P glycoprotein [42, 43]. Therefore, many attempts have been made to improve the properties of this natural compound with the goal to produce a more favorable therapeutic window. These efforts have resulted in three generations of epothilones under clinical development (Table 5).

First-generation epothilones are the compounds originally isolated from naturally occurring bacteria.

Second-generation epothilones are semisynthetic analogs of the natural products which have been modified to improve efficacy and tolerability.

Third-generation epothilones are fully synthetic and are designed to combine high efficacy, improved tolerability and lack of recognition by cellular resistance mechanisms in various solid tumors, including tumor types that are generally not sensitive to chemotherapy.

First results from clinical trials with some epothilones are promising indicating that epothilones have the potential to become a cornerstone in future cancer therapy.

6. References

General References

1. G. Höfle, H. Reichenbach: "Epothilone, a Myxobacterial Metabolite with Promising Antitumor Activity", in G. M. Cragg, D. G. I. Kingston, D. J. Newman (ed.): *Anticancer Agents From Natural Products*, CRC Press Taylor& Francis Group, Boca Raton 2005, pp. 413–450.
2. E. B. Watkins, A. G. Chittiboyina, Jae-Chul Jung; M. A. Avery: "The epothilones and related analogues — A review of their syntheses and anti-cancer activities", *Curr. Pharm. Design* **11** (2005) 1615.
3. D. Schinzer: "The aldol reaction in natural product synthesis. The epothilone story", in R. Mahrwald (ed.): *Modern Aldol Reactions*, Vol. 1, Wiley-VCH, Weinheim, Germany 2004, pp. 311–328.
4. K.-H. Altmann: "The merger of natural product synthesis and medicinal chemistry: on the chemistry and chemical biology of epothilones", *Org. Biomol. Chem.* **2** (2004) 2137.
5. R. Zhang, Z. Liu: "Recent progress in the synthesis of epothilones", *Curr. Org. Chem.* **8** (2004) 267.
6. K.-H. Altmann: "Epothilone B and its Analogs — A New Family of Anticancer Agents", *Mini Review in Medicinal Chemistry* **3** (2003) 149.
7. T. Beckers, S. Mahboobi: "Natural, semisynthetic and synthetic microtubule inhibitors for cancer therapy", *Drugs Future* **28** (2003) 767.
8. R. M. Borzilleri, G. D. Vite: "Epothilones: new tubulin polymerization agents in preclinical and clinical development", *Drugs Future* **27** (2002) 1149.
9. R. Altaha, T. Fojo, E. Reed, J. Abraham: "Epothilones: a novel class of non-taxane microtubule-stabilizing agents", *Curr. Pharm. Design* **8** (2002) 1707.
10. M. Wartmann, K.-H. Altmann: "The biology and medicinal chemistry of epothilones", *Curr. Med. Chem.: Anti-Cancer Agents* **2** (2002) 123.
11. M. A. Jordan: "Mechanism of action of antitumor drugs that interact with microtubules and tubulin", *Curr. Med. Chem.: Anti-Cancer Agents* **2** (2002) 1.
12. U. Klar, W. Skuballa, B. Buchmann, W. Schwede, T. Bunte, J. Hoffmann, R. B. Lichtner: "Synthesis and biological activity of epothilones", in I. Ojima, G. D. Vite, K.-H. Altmann (ed.): *Anticancer Agents. Frontiers in Cancer Chemotherapy*, Oxford University Press, New York 2001, 131–147.
13. M. Kavallaris, N. M. Verrills, B. T. Hill: "Anticancer therapy with novel tubulin-interacting drugs", *Australia Drug Resistance Updates* **4** (2001) 392.
14. K.-H. Altmann, G. Bold, G. Caravatti, N. End, A. Flörsheimer, V. Guagnano, T. O'Reilly, M. Wartmann: "Epothilones and their analogs — potential new weapons in the fight against cancer", *Chimia* **54** (2000) 612.
15. G. Appendino, G. Casiraghi: "The synthesis of epothilones: highlights from a year's race", *Chemtracts* **11** (1998) 678.
16. S. B. Buck, S. F. Victory, R. H. Himes, G. I. Georg: "Epothilones, a new class of microtubule-stabilizing agents with a

taxol-like mechanism of action", *Chemtracts* **11** (1998) 671.

Specific References

17. Gesellschaft für Biotechnologische Forschung mbH, DE 4138042 C2, 1991 (G. Hoefle, N. Bedorf, K. Gerth, H. Reichenbach)
18. D. M. Bollag et al.: "Epothilones, a New Class of Microtubule-Stabilizing Agents with a Taxol-like Mechanism of Action", *Cancer Res.* **55** (1995) 2325.
19. D. M. Bollag: "Epothilones: Novel Microtubule-Stabilizing Agents", *Expert Opin. Invest. Drugs* **6** (1997) 867.
20. O. A. O'Connor et al.: "Stability and toxicity of 12,13-Desoxyepothilone B (DEPOB) in dogs with spontaneous tumors", poster presentation 92th AACR Meeting, New Orleans, March 24-28 (2001).
21. G. Höfle et al.: "Epothilon A und B — neuartige, 16gliedrige Makrolide mit cytotoxischer Wirkung: Isolierung, Struktur im Kristall und Konformation in Lösung", *Angew. Chem.* **108** (1996) 1671, *Angew. Chem., Int. Ed.* **35** (1996) 1567.
22. R. E. Taylor; Y. Chen, A. Beatty, D. C. Myles, Y. Zhou: "Conformation–Activity Relationships in Polyketide Natural Products: A New Perspective on the Rational Design of Epothilone Analogues", *J. Am. Chem. Soc.* **125** (2003) 26.
23. D. Schinzer, A. Limberg, A. Bauer, O. M. Böhm, M. Cordes: "Totalsynthese von (−)-Epothilon A", *Angew. Chem.* **109** (1997) 543; *Angew. Chem., Int. Ed.* **36** (1997) 523.
24. D. Schinzer, A. Limberg, O. M. Böhm: "Studies Towards the Total Synthesis of Epothilones: Asymmetric Synthesis of the Key Fragments", *Chem. Eur. J.* **2** (1996) 1477.
25. A. Balog et al.: "Totalsynthese von (−)-Epothilon A", *Angew. Chem.* **108** (1996) 2976, *Angew. Chem., Int. Ed.* **35** (1996) 2801.
26. P. Bertinato, E. J. Sorensen, D. Meng, S. J. Danishefsky: "Studies towards a Synthesis of Epothilone A: Stereocontrolled Assembly of the Acyl Region and Models for Macrocyclization", *J. Org. Chem.* **61** (1996) 8000.
27. K. C. Nicolaou et al.: "Total Syntheses of Epothilones A and B via a Macrolactonization-Based Strategy", *J. Am. Chem. Soc.* **119** (1997) 7974.
28. D. Meng et al.: "Remote Effects in Macrolide Formation through Ring-Forming Olefin Metathesis: An application to the Synthesis of Fully Active Epothilone Congeners", *J. Am. Chem. Soc.* **119** (1997) 2733.
29. L. Tang et al.: "Cloning and heterologous expression of the epothilone gene clust", *Science* **287** (2000) 640.
30. I. Molnar et al.: "The biosynthetic gene cluster for the microtubule-stabilizing agents epothilones A and B from Sorangium cellulosum So ce90", *Chem. Biol.* **7** (2000) 97.
31. B. Julien et al.: "Isolation and characterization of the epothilone biosynthetic gene cluster from Sorangium cellulosum", *Gene* **249** (2000) 153.
32. E. B. Watkins, A. G. Chittiboyina, J.-C. Jung, M. A. Avery: "The epothilones and related analogues-A review of their syntheses and anti-cancer activities", *Curr. Pharm. Design* **11** (2005) 1615.
33. B. Julien, S. Shah: "Heterologous expression of epothilone biosynthetic genes in *Myxococcus xanthus*", *Antimicrob. Agents Chemother.* **46** (2002) 2772.
34. R. L. Arslanian et al.: "Large-Scale Ilaon and Crystallizaon of Epothilone D From *Myxococcus xanthus* Cultures", *J. Nat. Prod.* **64** (2002) 570.
35. J. Lau, S. Frykman, R. Regentin, S. Ou, H. Tsuruta, P. Licari: "Optimizing the heterologous production of epothilone D in *Myxococcus xanthus*", *Biotechnol. Bioeng.* **78** (2002) 280.
36. L. Melton: "Cellular pump changes prospect of disease resistance", *Lancet* **357** (2001) 1186.
37. D. M. Bollag et al.: "Epothilones, a New Class of Microtubule-Stabilizing Agents with a Taxol-like Mechanism of Action", *Cancer Res.* **55** (1995) 2325.
38. R. J. Kowalski, P. Giannakakou, E. Hamel: "Activities of Mictotubule-stabilizing Agents Epothilones A and B with Purified Tubulin and in Cells Resistant to Paclitaxel (Taxol®)", *J. Biol. Chem.* **272** (1997) 2534.
39. J. H. Nettles et al.: "The Binding Mode of Epothilone A on α, β-Tubulin by Electron Crystallography", *Science* **305** (2004) 866.
40. E. Nogales, S. G. Wolf, K. H. Downing: "Structure of the α, β-Tubulin dimer by electron crystallography", *Nature* **391** (1998) 199.
41. G. Vite: "Synthesis and biological activity of epothilones", In: *New Prospects in Anticancer Agents for the 21st Century*, American Chemical Society, 2001, pp. 131–147.

42. U. Klar et al.: "A new potent epothilone derivative superior to Epothilone B", *Proc. Am. Assoc. Cancer Res.* **43** (2002) 791, abs 3924.
43. R. B. Lichtner et al.: "Subcellular distribution of epothilones in human tumor cells", *J. Proc. Natl. Acad. Sci. USA* **98** (2001) 11743.
44. K.-H. Altmann, M. Wartmann, T. O'Reilly: "Epothilones and related structures — a new class of microtubule inhibitors with potent in vivo antitumor activity", *Biochim. Biophys. Acta* **1470** (2000) M79.
45. S. Mani et al.: "Phase I Clinical and Pharmacokinetic Study of BMS-247550, a Novel Derivative of Epothilone B, in Solid Tumors", *Clin. Cancer Res.* **10** (2004), 1289.

Microreactors – Modeling and Simulation

Steffen Hardt, TU Darmstadt, Darmstadt, Germany

Related Articles

1. Introduction 1
2. Flow Distributions 5
2.1. Straight Microchannels 5
2.2. Periodic and Curved Channel Geometries 6
2.3. Multichannel Flow Domains 7
3. Heat Transfer 11
3.1. Straight Microchannels 11
3.2. Periodic and Curved Channel Geometries 12
3.3. Multichannel Flow Domains 13
3.4. Micro Heat Exchangers 15
4. Mass Transfer and Mixing 18
4.1. Simple Mixing Channels 18
4.2. Chaotic Micromixers 20
4.3. Multilamination Micromixers 25
4.4. Hydrodynamic Dispersion 27
5. Chemical Kinetics 29
5.1. Numerical Methods for Reacting Flows 30
5.2. Reacting Channel Flows 31
5.3. Heat-Exchanger Reactors 33
6. Conclusion and Outlook 34
7. References 35

1. Introduction

In the field of conventional, macroscopic process technology modeling and simulation approaches are by now used on a routine basis to design and optimize processes and equipment. Many of the models employed have been developed for and carefully adjusted to specific processes and reactors and allow flow and heat and mass transfer to be predicted, often with a high degree of accuracy. In comparison, modeling and simulation approaches for microreactors are more immature, but have great potential for even more reliable computer-based process engineering. In general, the purposes of computer simulations are manifold, such as feasibility studies, optimization of process equipment, failure modeling, and modeling of process data. For each of these tasks within the field of chemical engineering simulation methods have been applied successfully.

Microreactors (→ Microreactors) are developed for a variety of different purposes, specifically for applications which require high heat- and mass-transfer coefficients and well-defined flow patterns. The spectrum of applications includes gas and liquid flow as well as gas – liquid or liquid – liquid multiphase flow. The variety and complexity of flow phenomena clearly pose major challenges to the modeling approaches, especially when additional effects such as mass transfer and chemical kinetics must be taken into account. However, there is one aspect which makes the modeling of microreactors in some sense much simpler than that of macroscopic equipment: the laminarity of the flow. Typically, in macroscopic reactors the conditions are such that a turbulent flow pattern develops, which makes the use of turbulence models [1] necessary. With turbulence models the stochastic velocity fluctuations below the scale of grid resolution are accounted for in an effective manner, without the need to explicitly model the time evolution of these fine details of the flow field. Heat- and mass-transfer processes strongly depend on the turbulent velocity fluctuations, and for this reason the accuracy of the turbulence model is of paramount importance for a reliable prediction of reactor performance. However, currently (2006) there is no model available which is capable of describing turbulent flow phenomena in a universal manner and is computationally inexpensive at the same time. For this reason, simulation approaches for microreactors, which usually do not require turbulence models, offer some potential to make predictions with a degree of accuracy unparalleled by models of macroscopic reactors.

When comparing processes in microreactors with those in conventional systems, a few general differences can be identified:

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.k16_k01

- Flow in microstructures is usually laminar, in contrast to the turbulent flow patterns on the macroscale.
- The diffusion paths for heat and mass transfer are very short, and this makes microreactors ideal candidates for heat- or mass-transfer-limited reactions.
- The surface-to-volume ratio of microstructures is very high. Thus, surface effects are likely to dominate over volumetric effects.
- The fraction of solid wall material is typically much higher than in macroscopic equipment. Thus, solid heat-transfer plays an important role and must be accounted for when designing microreactors.

While the absence of turbulence simplifies many modeling tasks, the predominance of surface effects introduces additional complications, especially in the case of multiphase flow. Some of the fundamental mechanisms, for example, those of dynamic wetting and spreading phenomena, are not yet well understood, and this adds some degree of uncertainty to the modeling of these processes. As more and more practical applications of microfluidic systems emerge, research in the field of fluidic surface and interfacial phenomena gains additional impetus. It is thus hoped that in the following years refined models for microfluidic multiphase systems will be formulated and will add an additional degree of predictability to flow phenomena in microreactors.

When facing the task of formulating a model of a specific microfluidic system, the question arises whether the conventional macroscopic equations describing fluid flow and heat and mass transfer are still valid on the microscale. Systems for chemical processing rarely contain structures with dimensions smaller than 10 μm; the relevant length scale is often in the range of 100 μm. The standard approach to modeling transport processes in microreactors relies on a continuum description, i.e., continuum-field quantities such as pressure and velocity are introduced to represent the mean dynamics of the fluid molecules. The corresponding mathematical formulation is based on the Navier – Stokes equation and convection – diffusion equations for heat and mass transfer. The question to answer is whether or not this mathematical framework is suitable to describe transport processes in microreactors on the relevant length scales. Besides the continuum assumption there is one additional assumption on which the Navier – Stokes and related equations rest. This is related to the local statistical distribution of the particles in phase space: The standard convection – diffusion equations rely on the assumption of local thermal equilibrium. For gas flow, this means that a Maxwell – Boltzmann distribution is assumed for the velocity of the particles in the frame of reference co-moving with the fluid. When gas flow in microreactors at high temperature or low pressure is considered, this assumption may break down. The principle quantity determining the flow regime of gases and deviations from the standard continuum description is the Knudsen number Kn (Eq. 1).

$$\mathrm{Kn} = \frac{\lambda}{L} \tag{1}$$

The Knudsen number is the ratio of two length scales: the mean free path of the gas molecules λ and a characteristic length scale of the flow domain L, for example, the channel diameter. When Kn is larger than unity, a gas molecule is more likely to collide with the channel wall than with another molecule. As the transport of momentum or enthalpy is to a large extent governed by the collisions between molecules, major changes in the flow behavior are expected when the Knudsen number is on the order of unity. This may happen when gas flow through narrow channels is considered, but also when the temperature is high and/or the pressure is low.

Based on the Knudsen number, four different flow regimes can be distinguished [2]:

- Continuum flow with no-slip boundary conditions ($Kn \leq 10^{-2}$)
- Continuum flow with slip boundary conditions ($10^{-2} < Kn \leq 10^{-1}$)
- Transition flow ($10^{-1} < Kn \leq 10$)
- Free molecular flow ($Kn > 10$)

In the first two cases the Navier – Stokes equation can be applied, with modified boundary conditions in the second case. The computationally most difficult case is the transition flow regime, which, however, may be encountered in microreactor systems. Clearly, the defined ranges of Knudsen numbers are not rigid; they instead vary from case to case. However, the numbers given above are guidelines applicable to many situations encountered in practice.

For applications in the field of microreaction engineering, the conclusion can be drawn that the Navier – Stokes equation and other continuum models are valid in many cases, as Knudsen numbers greater than 10^{-1} are rarely obtained. This observation is supported by the fact that a typical mean free path of nitrogen at 1 bar is about 70 nm. However, it may be necessary to use slip boundary conditions to describe the flow in microchannels. The first theoretical investigations on slip flow of gases were carried out already in the 1800s by MAXWELL and VON SMOLUCHOWSKI. The basic concept relies on a so-called slip length L_s, which relates the local shear strain to the relative flow velocity at the wall (Eq. 2)

$$u_{\mathrm{gas}} - u_{\mathrm{wall}} = L_{\mathrm{s}} \frac{\partial u_{\mathrm{gas}}}{\partial y} \Big|_{\mathrm{Wall}} \tag{2}$$

where u_{gas}, u_{wall} refer to the gas and wall velocities parallel to the wall and y is the coordinate normal to the wall. The slip length L_s is depends on wall roughness and can be determined experimentally. Similarly, there is a relationship for the temperature jump across a solid surface in the slip-flow regime [2].

In the transition flow regime, which is rarely encountered in microreactors, the usual convection – diffusion equations cease to be valid. The general formulation to be applied in this regime is given by the Boltzmann equation, from which less involved model descriptions such as the Burnett equation can be derived via a low-*Kn* expansion [3]. However, rather than by a continuum model, transition flows are often computed based on the Direct Simulation Monte Carlo (DSMC) method [4]. Instead of modeling gas flow by a set of differential or integro-differential equations, this method is based on tracking the trajectories and interactions of gas molecules directly. DSMC is a time marching approach based on a time discretization with steps smaller than the collision time of the molecules. The computational demand for DSMC is much higher than for continuum approaches.

Similar to the case of gas flows, the question of the range of applicability of continuum models for liquid flow arises. While the kinetic theory of gases provides clear indications of the limits of continuum models and of the onset of rarefaction effects on the microscale, there is no general framework explaining possible deviations of liquid flow phenomena from their macroscopic behavior. The concept of mean free path ceases to be useful for liquids, as the molecules interact with their neighbors in a permanent manner (in a sense, the molecular mean free path is zero for liquids). Some groups have conducted experiments on liquid flows in microstructures and measured pressure drop and heat-transfer coefficients, with largely contradictory results [5 – 11]. A compilation of the published results does not seem to leave room for any simple conclusions of general importance. Furthermore, the physics of Newtonian liquids does not suggest any universal mechanisms by which the flow behavior in channels of several hundred micrometers in width could differ considerably from macroscopic behavior at the same Reynolds number. However, in some specific situations liquid microflows exhibit a behavior which does not occur on the macroscale, but which can be reproduced experimentally and explained by theoretical models.

Boundary slip of liquids is a first example of such effects. Since ca. 1995 it was confirmed that not only gases, but also liquids can exhibit boundary slip, with corresponding experiments indicating a slip length of a few tens of nanometers [12 – 14]. The phenomenon has also been studied by numerical simulations, especially the molecular dynamics (MD) method, which is based on solving Newton's equations of motion for an ensemble of mutually interacting molecules (see, e.g., [15]). MD simulation results confirm the existence of boundary slip for liquids and indicate how the slip length depends on molecular parameters and shear rate [16].

While slip-flow effects are mostly expected to have an impact on flow behavior on the submicrometer scale, there is another phenomenon which may lead to deviations from the usual model predictions on a larger length scale. On a solid surface exposed to an ionic liquid, charges may accumulate to form a Debye layer above the surface, to which an additional layer of immobilized charges is attached [17]. This so-called electric double layer (EDL) is related to accumulation of charges inside the liquid which act as sources of an electric field. In turn, this field is sensitive to changes in the EDL structure which may be induced by the flow in such a manner that inside a channel a field component in streamwise direction builds up. When ionic charges start to migrate in the electric field, liquid molecules

are dragged along with them and the apparent viscosity for channel flow changes. The phenomenon is known as electroviscous effect [18] and may cause a considerable modification of the friction factor in comparatively large channels. The electroviscous effect was studied theoretically for rectangular microchannels [19]. The channel width and height considered were in the range between 20 and 40 μm. Depending on the solute concentration, the friction factor was increased by up to 30 %, which shows that the electroviscous effect manifests itself on length scales much larger than the EDL thickness, typically on the order of a few nanometers. Thus, even in typical microchannel geometries ionic liquids may exhibit some "unexpected" behavior.

Many of the flow phenomena discussed here can be incorporated into continuum models such as the Navier – Stokes equation, an exception being gas flows in the transition and free molecular regimes. Only a modification of the boundary conditions and not of the model equations themselves is usually necessary, i.e., the nonstandard physics characteristic for the micro- and nanoscale mainly appears as a boundary effect. Thus, when attempting to model transport processes in microreactors, the continuum hypothesis is a far-reaching concept which covers most of the scenarios encountered in practice. For a more detailed discussion of the special aspects of micro- and nanoflows, see [20]. Along this line of arguments it becomes clear that to a large extent the problem of modeling transport phenomena in microreactors is a scaling problem: The familiar macroscopic model equations may be used, but they must be applied in an unfamiliar and possibly unexplored parameter range as far as the dimensionless groups characterizing the problem are concerned [21]. In that context, care must be taken to pay attention to effects which are usually neglected macroscopically, such as surface tension and viscous heating.

Most of the simulation techniques and results presented in the following chapters are based on a numerical solution of the macroscopic transport equations. As already mentioned above, owing to the laminar flow conditions, many of these studies can be regarded as first-principles calculations which are free of adjustable model parameters. Consequently, simulation results for microreactors are expected to have a high predictive power. The major limitation for the predictive power of the models is the large number of degrees of freedom, which often excludes the simulation of complete reactors and only allows focusing on components such as single channels. Thus, the true art of reactor modeling consists of selecting a suitable subsystem and representing the portions of the reactor neglected by appropriate boundary conditions. For many of the results reported in the following sections such a strategy was employed. When it is not clear how to select an appropriate subsystem or genuine system-level behavior to be studied, a macromodel approach may be chosen. One strategy of doing so is based on partitioning of the reactor domain into certain sections, each of which is characterized by a specific input – output behavior. As input to the model sections, thermodynamic forces such as pressure or temperature gradients are applied by which thermodynamic fluxes such as mass or heat flux are induced. An alternative strategy of macromodel generation is based on an averaging approach. By averaging over a large number of microscopic structures (e.g., channels) effective equations for volume-averaged field quantities can be derived, thus allowing the number of degrees of freedom for system-level modeling to be reduced considerably. In the following chapters, some examples for both of these modeling strategies which allow transport phenomena to be studied on the level of complete systems or at least larger subsystems will be given.

With continuum models being the basis for the simulation of microreactors, computational fluid dynamics (CFD) is the numerical method to be applied. By now, CFD techniques are routinely used in conventional chemical process engineering. In principle the very same methods can be employed for microreactors, with the possible drawback that the numerical algorithms have been optimized mainly for the parameter regime characteristic for macroscopic flows. The fundamentals of CFD are discussed in → Computational Fluid Dynamics; suitable introductions can be found in, e.g., [22].

2. Flow Distributions

Due to the specific microstructuring technology employed to build up microreactors, the geometric shape of the flow domain is often different from that in macroscopic equipment. While usually the elements for fluid transport such as pipes are of circular cross section, the channels in microreactors have a rectangular or trapezoidal cross section. Depending on the specific microstructuring technology, also tub-like grooves or close-to-triangular shapes may be obtained. Furthermore, a characteristic feature common to many microreactors is a comparatively wide flow distributor followed by a large number of parallel microchannels. In the following, the flow distributions in characteristic geometries and methods to obtain approximate flow distributions in highly parallel flow domains are discussed.

2.1. Straight Microchannels

For laminar flow in channels of rectangular cross section, the velocity profile can be determined analytically. For this purpose, incompressible flow is assumed. The flow profile can be expressed in form of a series expansion (see [23] and references therein) which, however, is not always useful for practical applications, where often only a fair approximation of the velocity field over the channel cross section is needed. An approximate solution of the form [24]

$$u(x,y)=u_{\max}\left[1-\left(\frac{x}{a}\right)^s\right]\left[1-\left(\frac{y}{b}\right)^r\right] \tag{3}$$

was suggested for a rectangular channel oriented along the z-axis with a width of $2a$ in the x-direction and a depth of $2b$ in the y-direction, where u is the local flow velocity and $u_{\max}$ the maximum velocity. The exponents s and r depend on the aspect ratio b/a of the channel. The most common correlations can be found in [23]. Typically, Equation 3 approximates the exact velocity profile with an accuracy of a few percent. Often, not the detailed velocity profile is of interest, but only the friction factor f, which determines the pressure drop over a channel of a given length. The fanning friction factor is defined as

$$f=-\frac{\mathrm{d}p}{\mathrm{d}z}\frac{D_\mathrm{h}}{2\rho U^2} \tag{4}$$

where $\mathrm{d}p/\mathrm{d}z$ is the pressure gradient along the channel, ρ the density, D_h the hydraulic diameter and U the mean flow velocity. The hydraulic diameter is given by $4A/P$, where A is the cross-sectional area of the channel and P its perimeter. A comparatively simple expression for the friction factor in rectangular channels [25] deviates from the analytical solution by less than 0.05 % (Eq. 4)

$$f=\frac{24}{\mathrm{Re}}\left(1-1.3553\alpha+1.9467\alpha^2-1.7012\alpha^3+0.9564\alpha^4-0.2537\alpha^5\right) \tag{5}$$

where α is the aspect ratio of the channel ($\alpha \leq 1$ by definition) and the length scale entering the Reynolds number Re is the hydraulic diameter.

Most expressions for the flow profile in rectangular channels assume that the flow is fully developed, i.e., that the flow velocity is oriented along the z-axis and does not change in streamwise direction. Close to the entrance of a channel this assumption is not valid, and the flow undergoes a development from an entrance distribution to a fully developed profile. Correspondingly, the pressure distribution deviates from that observed in a fully developed flow, and expressions for the friction factor such as Equation 5 are not valid in the entrance region. To determine the developing velocity distribution and friction factor, various approaches have been employed, such as analytical calculations based on a linearized inertia term, numerical solutions of the Navier – Stokes equation, and experimental velocity and pressure measurements. An overview of the results is given in [23]. In general, the pressure drop per unit length of an entrance flow is higher than that of a fully developed flow. To compare entrance flow effects, a hydrodynamic entrance length L_hy can be defined which is the length necessary to achieve a centerline velocity equal to 99 % of the fully developed value. Usually, a nondimensional quantity is used (Eq. 6).

$$L_\mathrm{hy}^+=\frac{L_\mathrm{hy}}{D_\mathrm{h}\mathrm{Re}} \tag{6}$$

Depending on the aspect ratio of the channel, values between 0.01 and 0.1 are found for the nondimensional entrance length [23]. From Equation 6 it can be deduced — given that L_hy^+ is Re-independent — that L_hy increases linearly

with the hydraulic diameter and the Reynolds number.

A number of authors have considered channel cross sections other than rectangular [25 – 27]. In general, an analytical solution of the Navier – Stokes and enthalpy equations in such channel geometries would be quite involved due to the implementation of the wall boundary condition. For this reason, usually numerical methods are employed to study laminar flow and heat transfer in channels with arbitrary cross-sectional geometry.

2.2. Periodic and Curved Channel Geometries

A key advantage of microreactors is their potential for rapid heat and mass transfer. Naturally, the acceleration of transport processes is related to the decrease of diffusion paths on the microscale. In addition, special channel geometries are explored for which a further speed up of heat and mass transfer is found. Three different channel shapes that have been studied are shown in Figure 1: zigzag sinusoidally curved, and converging-diverging. Velocity and temperature fields in zigzag channels were computed based on the finite-volume method for Reynolds numbers up to 1500 [28]. Above a specific Reynolds number depending on the geometry parameters, flow separation occurs, i.e., in the corners of the channel recirculation zones are formed. While the friction factor for straight channels displays a linear decrease as a function of Reynolds number, as is apparent from Equation 5, the friction factor in zigzag channels becomes nearly independent of the Reynolds number for $Re > 1000$. Data obtained for a zigzag channel with rounded corners show a similar trend [29].

Flow in sinusoidally curved channels for Reynolds numbers up to 500 was also studied with a finite-volume discretization [30]. For most of the parameter space explored, flow separation was not observed. However, for $Re = 500$ and a large enough sine-wave amplitude as compared to the period, there were some indications of recirculation zones forming in the recesses of the channel. Unfortunately, no attempt was made to compute friction factors.

Flow in converging-diverging channels was studied with special attention to the transition between laminar and turbulent flow [31] by using a spectral-element method in which the flow domain is divided into macroelements over which a set of polynomial test functions is defined. This method allowed damping of small-scale fluctuations due to numerical viscosity to be suppressed and is thus well suited to studying the transition from stationary to oscillatory and chaotic flow. The computed streamline patterns indicate a transition from unseparated flow at low Reynolds numbers to a flow with recirculation zones within the recesses of the channel, occurring at a comparatively small Reynolds number between 10 and 20. Streamline patterns for various Reynolds numbers are depicted in Figure 2. The phenomenology of flow distributions in such channel domains was found to be quite diverse. At low Reynolds numbers the flow is stationary, but at a Reynolds number of 150 oscillations begin to develop, with the vortices still being confined in the recesses. At Re = 400 the viscous forces are no longer strong enough to confine the vortices in the recesses and vortex ejection is observed. At even higher Reynolds numbers the flow becomes aperiodic and chaotic.

In some cases, for example, when long residence times are desired, comparatively long microchannels must be integrated into a compact micro reaction device, a task which can only be achieved with curved channels. In addition, straight channels are often not suited for connecting a microflow domain to the external world. One of the first theoretical studies on flow in curved channels [32, 33] investigated the secondary flow perpendicular to the main flow direction induced by inertial forces by means of a perturbative analysis. The dimensionless group characterizing such transverse flows is the Dean number.

$$K = \mathrm{Re}\sqrt{\frac{D_\mathrm{h}}{R}} \tag{7}$$

where D_h is the hydraulic diameter and R the mean radius of curvature of the channel. The typical secondary flow pattern in curved channels is given by two counterrotating vortices separated by the plane of curvature. The strength of these vortices increases with increasing Dean number.

Originally, the studies of secondary flow in curved channels were performed for pipes of circular cross section. Of much greater relevance to

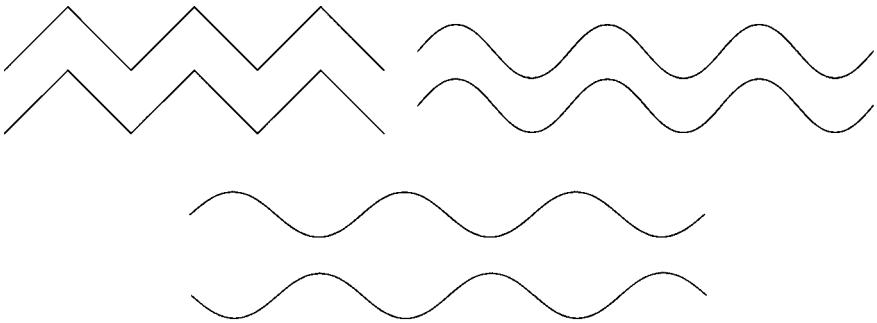

Figure 1. Different channel shapes for which flow distributions have been computed: zigzag (upper left), sinusoidally curved (upper right), and converging-diverging (bottom)

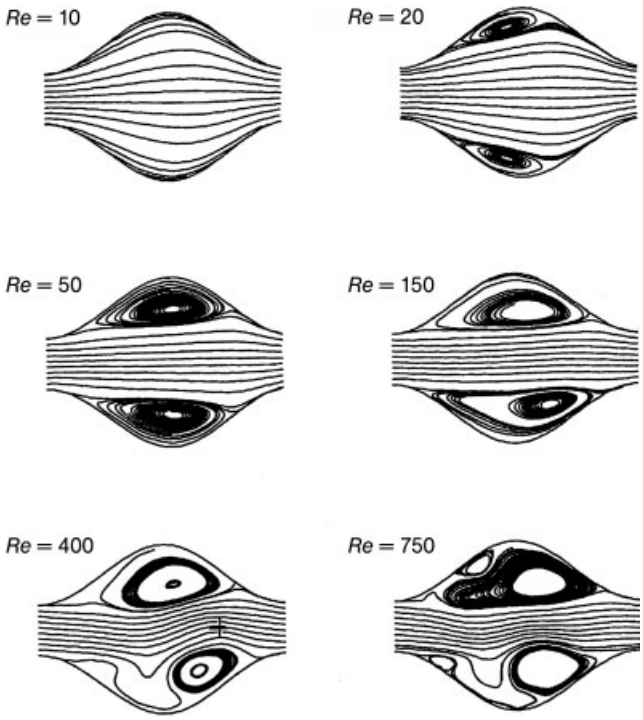

Figure 2. Streamline patterns in a converging-diverging channel for various Reynolds numbers, taken from [31]

microreactors are rectangular channels, which were the subject of detailed numerical investigations [34] into the formation of vortices in channels of square cross section, for which a complex branch structure was found describing the transition between flow patterns with two and four counterrotating vortices. Up to a Dean number of about 113, a two-vortex solution is found, followed by a transition regime between Dean numbers of 113 and 130 with coexisting two- and four-vortex configurations. For $K > 130$ the flow exhibits a four-vortex pattern which, however, coexists with a two-vortex pattern of broken symmetry. The friction factor in curved channels was found to be larger than in straight channels for most of the Dean number range considered, and not surprisingly different branches of the flow map are associated with different friction values.

2.3. Multichannel Flow Domains

In chemical microprocess technology it is important to guarantee well-defined and repro-

ducible reaction conditions in microchannels, but a high throughput should also be achieved. For this purpose, the process fluid is guided through a large number of parallel microchannels, where heat exchange and/or chemical reactions occur. One of the characteristic problems of microprocess technology is to equally distribute the incoming fluid over the microchannels. Fluid maldistribution would induce unequal residence times in different channels, with undesired consequences for the product distribution of a chemical reaction being conducted inside the reactor. When a process with various competing side reactions and byproducts is considered, the contact time of the process fluid in the reaction region should be as well defined as possible to maximize the selectivity of the process. However, not only the maldistribution of the fluid over a multitude of microchannels induces variations of the contact time, but also hydrodynamic dispersion of concentration tracers in the channels themselves.

Various concepts for the equipartition of fluid over a multitude of microchannels have been developed. One concept relies on guiding the incoming fluid through a flow splitter with subsequent bifurcations [35]. The most widely used design is based on a comparatively wide inlet region leading into a multitude of narrow microchannels. A corresponding geometry, developed for methanol reforming, is shown in Figure 3. The figure shows the computational domain of the CFD model of the reactor together with the computed streamlines, where use of reflection symmetry was made and only one quarter of the portion of the reactor to be considered was modeled. An inlet pipe leads to a flow distribution chamber connected with a multitude of microchannels. The essence of such designs is to be seen in the pressure barrier of the microchannels. The narrower the microchannels are, the higher will be the pressure drop in the channels themselves as compared to the pressure drop in the flow distribution chamber, and the more uniform the flow distribution will be.

During the development of the microreformer of Figure 3 one of the goals was to design the flow manifold in such a way that the volume flows in the different reaction channels are approximately the same [36]. In spite of the recirculation zones found, for the chosen design a flow variation of about 2 % between different channels was predicted from CFD simulations. In the application under study, a washcoat catalyst layer is applied to the microchannels. Thickness variations of the catalyst layer are likely to play the leading role in the nonuniformities in flow distribution to be expected. Hence, a flow manifold with intrinsic volume flow variations of a few percent over different reaction channels is usually satisfactory.

For the computation of the flow distribution in the methanol-reforming reactor, a reduced-order flow model for the microchannels was used. In such a model a fully developed flow profile as given by Equation 3 is assumed, which means that entrance flow effects are neglected. This approximation is justified when the entrance length is small compared to the total length of the channel. A major advantage of a reduced-order flow model is the significant reduction of the degrees of freedom entering the simulation. Each microchannel is then only represented by a single degree of freedom, that is, the total volume flow, and resolution of the corresponding flow domain by a computational grid is no longer required. Thus the simulation of microfluidic devices comprising a large number of channels becomes possible at moderate computational cost.

Finite-volume grids in combination with reduced-order flow models for microchannels allow flow distribution problems to be solved in 3D, for which the standard approach would be computationally too expensive. However, when the goal is to find designs with optimum flow equipartition by tuning specific geometric parameters, it is advisable to set up models with significantly fewer degrees of freedom. The flow distribution in multichannel microreactors was studied with the help of macromodels [37]. The geometry considered was a microstructured plate of a heat-exchanger stack developed at the Institute of Microtechnology Mainz. The plate together with arrows indicating the flow direction is shown on the left side of Figure 4; on the right side the model is displayed together with the geometric parameters. The model is based on the idea of subdividing the flow domain into a number of virtual channel segments with rectangular cross section over which the flow is distributed. Through the channel segments in the inlet zone the volume flows $V_1, V_2, \ldots, V_{\mathrm{Nc}}$ are transported, where a part of the flow branches off into the actual microchannels of width W_{c}

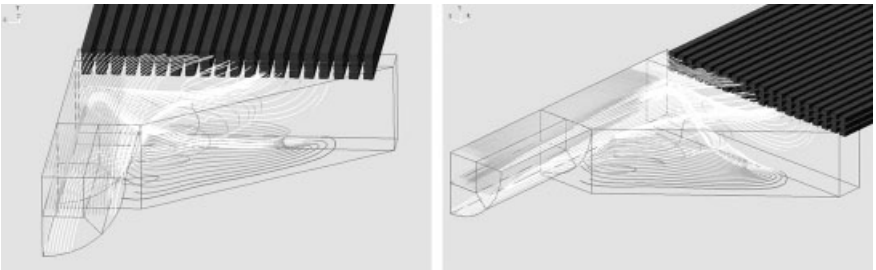

Figure 3. Streamlines in the flow distribution chamber of a multichannel reactor

and depth e. For each of the channel segments a relationship between the pressure drop Δp and the average flow velocity u of the form

$$\Delta p = 32\lambda_{\mathrm{Nc}} \frac{\mu L u}{D_{\mathrm{h}}^2} \tag{8}$$

was used, where the hydraulic diameter is defined as

$$D_{\mathrm{h}} = \frac{2we}{w+e} \tag{9}$$

In these expressions, μ is the dynamic viscosity, L the length of the channel segment, w and e are its width and depth, and λ_{Nc} is a correction factor accounting for the noncircularity of the channels. Clearly, the above formulas rely on the assumption of a hydrodynamically developed flow.

By means of the model depicted in Figure 4, Equation 8, and Equation 9 it is possible to compute the flow distribution just by solving a comparatively low-dimensional system of linear algebraic equations. The problem resembles the task of computing the current distribution of an electric circuit. A priori it is not clear whether the approximation to subdivide the flow domain in the described way is justified. To assess the quality of the chosen approximation, the flow field was computed by means of the finite-volume method [37]. The results obtained suggest that, owing to the orientation of the isobars of the flow and the absence of recirculation loops, the chosen subdivision into channel segments is a reasonable simplification, at least for the geometry and flow regime considered. The model allows the flow distribution to be studied for a variety of different geometries with a minimum computational effort. As a result of simulations, optimized flow chamber geometries leading to a maximum of flow equipartition were determined.

While the strategy described above is a way to obtain effective, low-dimensional models for multichannel reactors in various cases, it clearly has its limitations. In geometries with hundreds of microchannels, another method can be utilized to compute the flow field. In a situation where the flow is distributed over a large number of parallel microchannels, the multitude of channels can be regarded as a porous medium. For porous media effective, volume-averaged transport equations have been known for a long time (for an overview, see [38]). To solve a flow distribution problem for a multichannel geometry, the flow domain can be split into several regions, one of which is the flow distribution chamber, and another the region comprising the multitude of channels. In the flow distribution chamber, the ordinary transport equations are solved, whereas in the multichannel domain, the effective, volume-averaged description of the transport processes is used. A corresponding 2D geometry is shown in Figure 5. The incoming flow is distributed over a large number of channels of width of w_{c} separated by walls with a width of $w - w_{\mathrm{c}}$.

The most straightforward porous-medium model which can be used to describe the flow in the multichannel domain is the Darcy equation [38], which represents a simple model used to relate the pressure drop and the flow velocity inside a porous medium. Applied to the geometry of Figure 5 it is written as

$$\frac{\mathrm{d}}{\mathrm{d}y} \langle p \rangle_{\mathrm{f}} + \varepsilon \frac{\mu}{K} \langle u \rangle_{\mathrm{f}} = 0 \tag{10}$$

with the porosity ε and permeability K given as

$$\varepsilon = \frac{w_{\mathrm{c}}}{w}, \quad K = \frac{\varepsilon\, w_{\mathrm{c}}^2}{12} \tag{11}$$

The flow velocity, pressure and dynamic viscosity are respectively denoted as u, p, and μ and

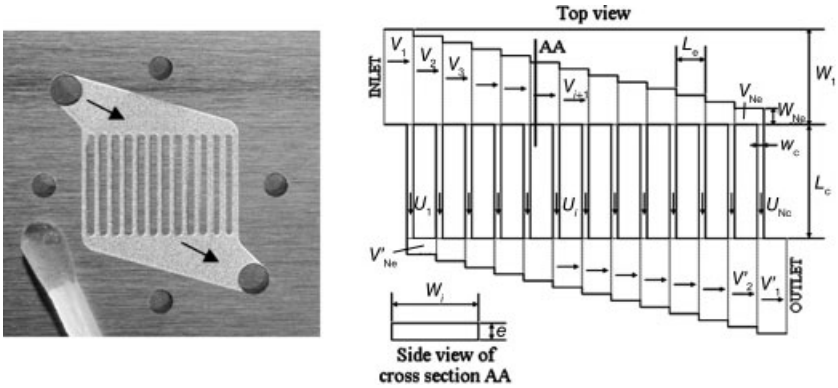

Figure 4. Geometry of the microstructured plates (left) and subdivision of the flow domain into channel segments (right), as considered in [37]

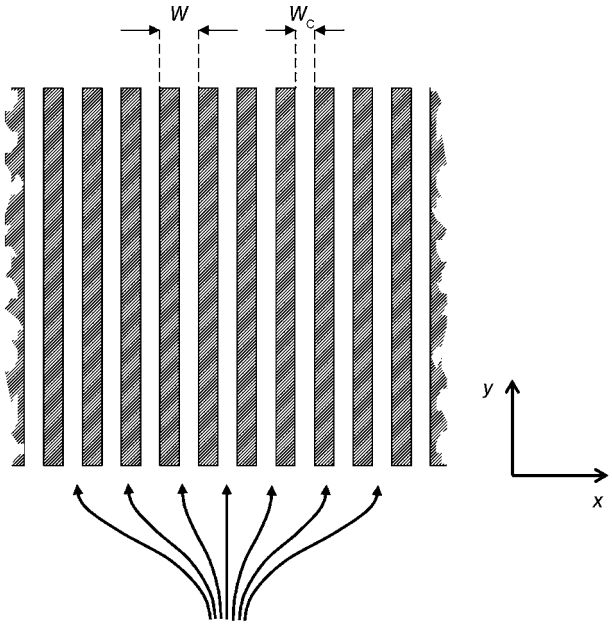

Figure 5. Multichannel geometry with channels and separating walls of uniform width

the symbol $<\ldots>_{\mathrm{f}}$ represents an average over the fluid phase. [39]. An extended Darcy equation was used to model the flow distribution in a microchannel cooling device. In general, the permeability K must be regarded as a tensor quantity accounting for the anisotropy of the medium. Furthermore, the description can be generalized to include heat-transfer effects in porous media.

Compared to the use of reduced-order flow models, the porous-medium approach allows an even larger multitude of microchannels to be dealt with. Furthermore, for comparatively simple geometries with only a limited number of channels, it represents a simple way to provide qualitative estimates of the flow distribution. However, as a course-grained description it does not reach the same level of accuracy as reduced-order models. Compared to the macromodel approach [37], the porous-medium approach has a broader scope of applicability and can also be applied when recirculation zones appear in the flow distribution chamber. However, the macromodel approach is computationally less expensive and can ideally be used for optimization studies.

3. Heat Transfer

Heat-transfer phenomena belong to the key issues to be studied in microreactors. Due to the small thermal diffusion paths, microreactors have the potential to enable fast heat transfer and to control temperature distributions with very high accuracy. Correspondingly, modeling and simulation of heat-transfer phenomena and their reliable prediction is of paramount importance for process design. The simulation of temperature distributions in microreactors often requires the solution of a conjugate heat transfer problem, i.e., the temperature fields in the fluid phase and in the solid wall material must be computed. Owing to the microstructuring technologies used to fabricate microreactors, the share of wall material in the total reactor volume is often higher than in conventional equipment. Hence, solid heat conduction effects become important and usually have to be taken into account when the temperature field inside a reactor is to be computed.

3.1. Straight Microchannels

A standard geometry of channels in microfluidic systems is a rectangular or close-to-rectangular cross section. With a given velocity profile, the temperature field inside rectangular channels can often be determined analytically. However, the problem of determining a temperature profile in a channel geometry is much more multifaceted than the computation of a flow distribution. While in most cases a zero-velocity boundary condition at the channel walls is prescribed for the flow, the wall-boundary conditions for the temperature field can be diverse. On the one hand, either a heat flux or a temperature can be prescribed. On the other hand, the thermal boundary conditions on the four walls of a rectangular channel may be different. Due to this complexity, the solution for temperature field itself is usually not reported. Rather, the Nusselt number, defined as

$$\mathrm{Nu}=\frac{\alpha D_{\mathrm{h}}}{k} \tag{12}$$

is determined. In this expression, D_{h} is the hydraulic diameter, k the thermal conductivity of the fluid, and α the heat-transfer coefficient measuring the transmitted thermal power per unit area divided by a characteristic temperature difference. The Nusselt number is a dimensionless quantity characterizing the efficiency of heat transfer. Similar to the velocity field, the temperature field assumes an invariant profile far enough downstream from the channel entrance. However, due to the continuous heat transfer from or to the channel walls, only the shape of the temperature distribution stays invariant, but the normalization changes. Close to the entrance of the channel a thermally developing flow may be observed. A thermal entrance flow is not a priori related to a hydrodynamic entrance flow, i.e., a thermally developing flow may be observed even in regions where the flow is hydrodynamically developed and vice versa.

An overview of the Nusselt numbers obtained for rectangular channels with various aspect ratios and various thermal boundary conditions has been given [23]. Depending on the thermal boundary conditions, the Nusselt number for thermally fully developed flow either increases or decreases when the aspect ratio is increased. When the Nusselt number for thermally developing and hydrodynamically developed flow is plotted as a function of position along the channel, a divergence is observed when the channel entrance is approached. This means that, with decreasing distance to the channel entrance, increasing heat transfer efficiency is found. The same observation is made when a simultaneously (i.e. hydrodynamically and thermally) developing flow is considered. Analogously to hydrodynamically developing flows, a thermal entrance length L_{th} can be defined. It is given as the duct length required for the Nusselt number to fall within a 5 % interval of the fully developed value. Again, a dimensionless quantity

$$L_{\mathrm{th}}^{+}=\frac{L_{\mathrm{th}}}{D_{\mathrm{h}}\mathrm{Re}\,\mathrm{Pr}} \tag{13}$$

is used, where the Prandtl number Pr is the ratio of the momentum diffusivity (i.e. the kinematic viscosity) and the thermal diffusivity. The dimensionless thermal entrance length is a quantity only depending on the aspect ratio of the channel and the thermal boundary conditions. Hence, L_{th} is a linear function of the hydraulic diameter, the Reynolds number, and the Prandtl number.

Apart from rectangular channels, heat transfer has also been studied in channels with different cross-sectional geometries, and Nusselt

numbers for wall boundary conditions of fixed temperature and fixed heat flux are tabulated for different geometric parameters [25, 26].

3.2. Periodic and Curved Channel Geometries

Similar to the computation of friction factors, heat transfer has not only been studied in straight channels, but also in channels with specific periodic shapes. Nusselt numbers in the zigzag channels shown in Figure 1 were computed by using a finite-volume approach in two dimensions [28]. With increasing Reynolds number the Nusselt number increases significantly as compared to a straight channel, i.e., a parallel-plate arrangement. For suitable geometric parameters, a Reynolds number in the range of 1000 and a Prandtl number of 8, a heat-transfer enhancement factor of about 13 was computed. For a Prandtl number of 0.7 (corresponding to air) the ratio of the heat fluxes achieved with the zigzag and the straight channels at identical values of pumping power and pressure drop were compared. The zigzag channel outperforms the straight channel by up to a factor of 6, i.e., for the same pumping power or pressure drop a considerably higher heat flux is achieved. This is an interesting result in view of the design of micro heat exchangers, in which the goal is often to increase the heat flux while limiting the pressure drop.

A zigzag geometry with rounded corners was studied in 2D by the finite-volume method [40]. The Prandtl number was kept fixed at 0.7 and the Reynolds number was varied between 100 and 1000. For the range of geometric parameters considered, a heat-transfer enhancement factor of about 3 was found at $Re = 1000$, and this confirms the potential of corrugated channels for heat-transfer enhancement.

Heat transfer in sinusoidally curved channels was studied for a Prandtl number of 1.0. [30] Surprisingly, at a Reynolds number of 100 the Nusselt number is lower than the corresponding value in a parallel-plate channel. Such a suppression of heat transfer stands in contradiction to results for sine-wave channel walls of comparatively large amplitude [41], for which heat-transfer enhancement increased with increasing Reynolds number, with an enhancement factor of about 4 for $Re = 800$. Both of these studies were restricted to steady-state flows, and steady-state solutions could be determined for the range of Reynolds numbers considered. Experimental work on flow and heat transfer in sinusoidally curved channels [42] indicate heat-transfer enhancement and do not show evidence of decreasing Nusselt number in any range of Reynolds numbers. However, the flow patterns begin to exhibit some unsteady behavior for Reynolds numbers greater than about 200 in the geometry considered. Interestingly, oscillations were predominantly observed close to the exit of the channel and were found to move upstream when the Reynolds number was increased. Much of the heat-transfer enhancement was attributed to the unsteady character of the flow [42].

The flow and temperature field in an elementary cell of the converging-diverging channel geometry for which flow fields were calculated in [31] was computed on the basis of the finite-volume method using periodic boundary conditions [43]. As far as the transition between a steady-state and an unsteady flow is concerned, similar results were found. At Reynolds numbers beyond the transition to the unsteady flow regime, the temperature field evaluated for a Prandtl number of 0.7 becomes quite complex. This is indicated in Figure 6, which shows the evolution of (normalized) temperature contour lines at a Reynolds number of 328. To quantify heat-transfer enhancement, the time-averaged Nusselt number was computed. The increase in Nusselt number compared to a straight channel was found to be significant and reached values as high as 7.54 at a Reynolds number of 520. Flow and heat transfer in converging-diverging channels were analyzed based on a perturbation analysis derived from lubrication theory and on the finite-volume method [44]. Most of these studies were limited to Reynolds numbers up to 100. In this regime the increase in Nusselt number compared to a parallel-plate geometry was moderate, amounting to a factor of about 1.3 at $Re = 100$ and $Pr = 1$. The results also showed an approximately linear increase of heat-transfer enhancement as a function of Prandtl number. The studies of [45] indicated that heat-transfer enhancement in converging-diverging channels is significantly reduced if the separation between the wavy channel walls is bigger than the wavelength of the undulations. Here a cubic

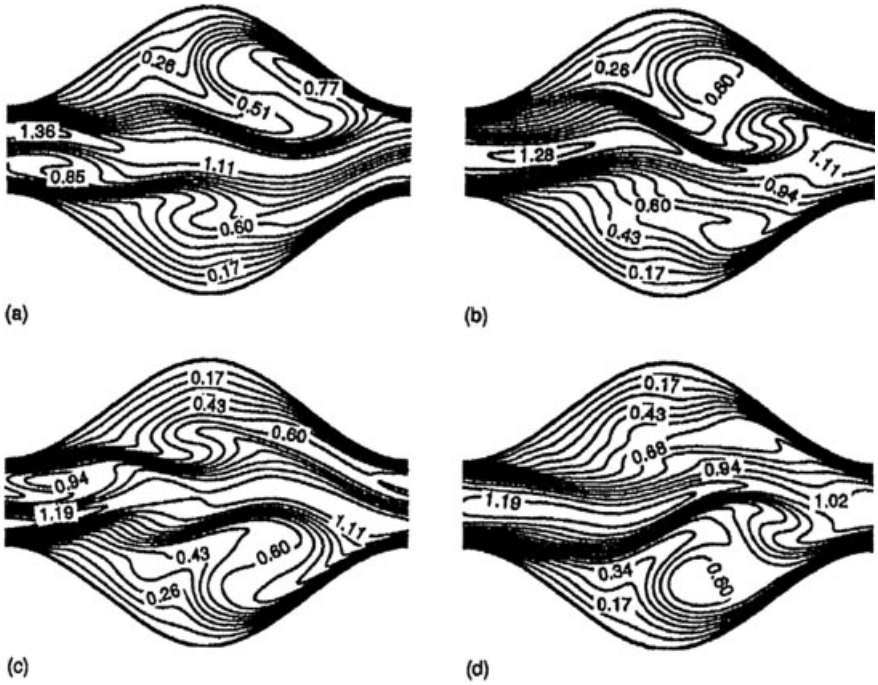

Figure 6. Evolution of the temperature field in the recesses of a converging-diverging channel, as obtained in [43]

spline collocation method was used to numerically solve the transport equations and, in contrast to the other studies, a cyclic structure of the flow and dimensionless temperature fields with the period of the wall structures was not assumed. The results obtained at $Re = 500$ and $Pr = 6.93$ show a rather moderate heat-transfer enhancement factor of about 1.8 compared to a parallel-plate geometry, which is further reduced for smaller Prandtl numbers.

Similar to the structure of the flow field, heat transfer has also been studied in curved channel geometries. The complicated branch structure with competing patterns of two and four counterrotating vortices in channels of square cross section is reflected in the Nusselt number [34]. When plotting the Nusselt number as a function of Dean number, different branches are found corresponding to symmetric and asymmetric secondary flow patterns with two and four vortices. However, the relative difference between the different branches is not very pronounced and should be hard to measure experimentally. For a Dean number of 210 and a Prandtl number of 0.7 a heat-transfer enhancement factor of about 2.8 was determined, that is, curved channels as well as other channels with specific periodically varying cross sections may be used for applications where rapid heat transfer is desired.

3.3. Multichannel Flow Domains

When a large number of parallel microchannels is considered, the problem of computing the temperature distribution inside the channels and the channel walls becomes quite involved. In such a case the use of the method of reduced-order flow models discussed in Section 2.3 is not as straightforward as in the case of a pure flow distribution problem, due to thermal cross-talk between the different channels. In principle, heat transfer in a multichannel domain can be described on the basis of a porous-medium model, in which an average is taken over an ensemble of channels and continuous field quantities are introduced. However, in the standard models for porous media, a thermal equilibrium between the fluid and the solid phase is assumed, i.e., only a single temperature field is used. Especially in the regime of large Péclet numbers the thermal-equilibrium assumption may break down, and it

may become necessary to define separate temperature fields for the fluid and the solid phase. It is thus advisable to take one step back and to derive the mean-field model equations that allow heat transfer in multichannel domains to be described on the basis of a volume-averaging approach.

A generic geometry of a multichannel stack to which the mean-field equations should be applied is shown in Figure 7. Many microchannels, usually several hundred up to several thousand, are arranged in layers stacked onto each other. As a whole the multichannel stack forms a cuboid of extension L_x, L_y, L_z in x-, y-, and z-direction, respectively. The microchannels of width w_f and depth h_f are assumed to be parallel to the z-axis and are separated by solid walls of thickness w_s and height h_s, respectively.

The need for modeling single channels and their specific geometry is eliminated when regarding the reaction fluid and wall material as interpenetrating continua. Hence, in the following it is assumed that the reactor stack is filled with interpenetrating fluid and solid phases which interact via the exchange of heat. Momentum exchange between the two phases is not accounted for explicitly and it is assumed that the velocity distribution over the cross section of the reactor is known a priori. The phase volume fractions can be derived from the geometric parameters h_f, h_s, w_f, and w_s and are given as

$$\Phi_\mathrm{f}=\frac{w_\mathrm{f}h_\mathrm{f}}{(w_\mathrm{f}+w_\mathrm{s})\,(h_\mathrm{f}+h_\mathrm{s})},\quad \Phi_\mathrm{s}=1-\Phi_\mathrm{f} \tag{14}$$

The temperature field of the solid phase can change due to heat conduction or due to heat transfer from the fluid phase. For the fluid, heat conduction does not have to be taken into account since the flow domain consists of disconnected channels, but convective heat transfer due to the flow velocity as well as heat transfer from the solid is included. In the xy plane no conduction within the fluid without intermediate transfer to the solid walls can occur. The reason for not including heat conduction in the z-direction is the fact that in most cases of practical relevance it is negligible when compared to convective transport. Following the derivation of [46], the fluid and solid temperature fields T_f and T_s are obtained as solutions of the equations

$$\begin{aligned}\rho_\mathrm{f}c_\mathrm{f}\left(\frac{\partial}{\partial t}+u_i\frac{\partial}{\partial x_i}\right)T_\mathrm{f}&=-\alpha a_\mathrm{V}\,(T_\mathrm{f}-T_\mathrm{s})+S_\mathrm{c}\\ \rho_\mathrm{s}c_\mathrm{s}\frac{\partial}{\partial t}T_\mathrm{s}&=k_{ij}\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}T_\mathrm{s}-\alpha a_\mathrm{V}\,(T_\mathrm{s}-T_\mathrm{f})\end{aligned} \tag{15}$$

where c_f, c_s are the fluid and solid specific heats, u_i is the average velocity of the fluid (averaged over a channel cross section), α the heat transfer coefficient from the channel walls to the fluid, a_V the specific surface area (wall surface/reactor volume), and the Einstein convention of summation over repeated indices has been used. This convention will be used throughout this chapter when vector or tensor quantities are involved. The fluid and solid densities are volume-averaged quantities and given as

$$\rho_\mathrm{f}=\Phi_\mathrm{f}\rho_\mathrm{f}^{(0)},\quad \rho_\mathrm{s}=\Phi_\mathrm{s}\rho_\mathrm{s}^{(0)}, \tag{16}$$

where the superscript (0) indicates the corresponding material densities. The equation for the fluid temperature contains a source term S_c which could, e.g., represent the input of thermal energy from a chemical reaction. Owing to the anisotropy of the solid matrix the thermal conductivity is direction-dependent. For this reason a thermal conductivity tensor k_ij appears in Equation 15. The heat-transfer coefficient α can be obtained from well-known correlations for rectangular channels. The model of Equation 15 can be regarded as a mean-field model expected to describe the average temperatures of the fluid and the solid phase without incorporating the local fluctuations which are due to temperature gradients within single channels or channel walls.

To compute the thermal conductivity tensor k_ij, the multitude of walls shown in Figure 7 can be regarded as a network of thermal resistors. On this basis, the components of the thermal conductivity tensor are derived as

$$\begin{aligned}k_{xx}&\approx(w_\mathrm{f}+w_\mathrm{s})\left(\frac{h_\mathrm{f}+h_\mathrm{s}}{h_\mathrm{s}}w_\mathrm{f}+w_\mathrm{s}\right)^{-1}k_\mathrm{s}\\ k_{yy}&\approx(h_\mathrm{f}+h_\mathrm{s})\left(\frac{w_\mathrm{f}+w_\mathrm{s}}{w_\mathrm{s}}h_\mathrm{f}+h_\mathrm{s}\right)^{-1}k_\mathrm{s}\\ k_{zz}&\approx\Phi_\mathrm{s}k_\mathrm{s}\end{aligned} \tag{17}$$

where k_s represents the thermal conductivity of the solid material. All other components of the thermal conductivity tensor vanish.

The mean-field model described here was developed to compute temperature fields in microreactors with low computational effort. Previously, a 2D version of such a model was used to determine the temperature distribution in a microchannel heat sink [39]. To assess the quality of the 3D mean-field model, the resulting

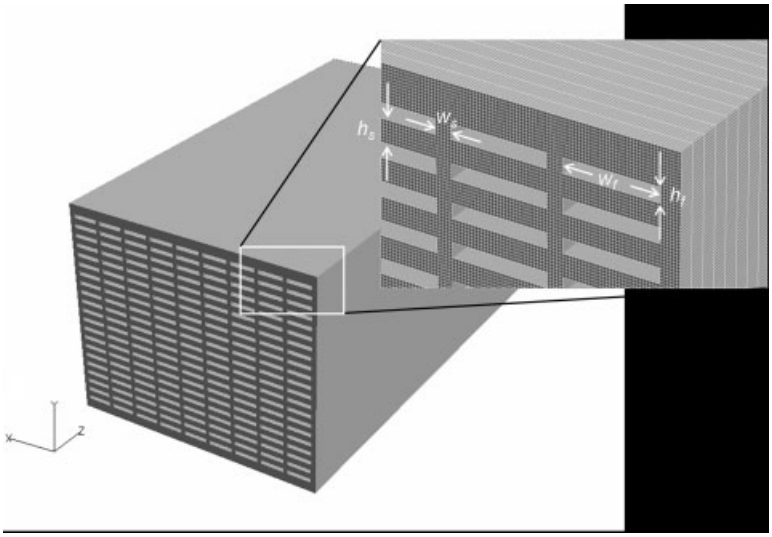

Figure 7. Multichannel stack comprising a large number of parallel microchannels

temperature fields were compared with results from a full conjugate heat-transfer model explicitly accounting for all of the geometrical details of the multichannel stack. The results of such a comparison are displayed in Figure 8. The temperatures were evaluated along the centerline of the reactor pointing in the z-direction. The heated region with nonzero source term extends up to 1 mm downstream from the inlet plane. The temperature curves show that in the inlet region the fluid and solid temperatures are not equilibrated, and this underlines the necessity of a two-temperature model. The same conclusion was drawn on the basis of the 2D model [39]. Overall, a good agreement between the mean-field and the full model is found, with maximum local differences in temperature of 2 – 3 K and a total temperature range of about 35 K.

The times needed to numerically solve the equations of the two models were found to differ by orders of magnitude. This makes the mean-field description an especially promising method for fast evaluation of reactor designs. When developing a micro reaction system it is often unclear how exactly the properties of the channel walls (geometry, thermal conductivity) influence the reaction performance. Usually, the goal is to achieve a temperature distribution as uniform as possible, in order to suppress unwanted side reactions and to increase the selectivity of the process. Based on the volume-averaging approach presented here it should be easy to assess and compare the thermal performance of different reactors and to identify favorable designs.

3.4. Micro Heat Exchangers

Micro heat exchangers are used for rapid heat transfer between a hot and a cold fluid; owing to the small thermal diffusion paths, the size of the system can be reduced compared to conventional devices. Among the different flow schemes (cocurrent, countercurrent, and cross-current) especially the countercurrent scheme has been studied theoretically, because of the practical relevance of countercurrent heat exchangers on the one hand, and the fairly simple model structure compared to cross-current heat exchangers on the other. A schematic drawing of a countercurrent micro heat exchanger is shown in Figure 9. A characteristic feature of such devices is the fact that the thickness of the wall material separating the channels is of the same order of magnitude as the channel depth itself. This distinguishes micro heat exchangers from their macroscopic counterparts, in which the width of the flow passage is usually well above the wall thickness.

Owing to the comparatively large share of wall material, longitudinal heat conduction in the channel walls must be taken into account when attempting to formulate a heat-exchanger model. In contrast, the models for conventional

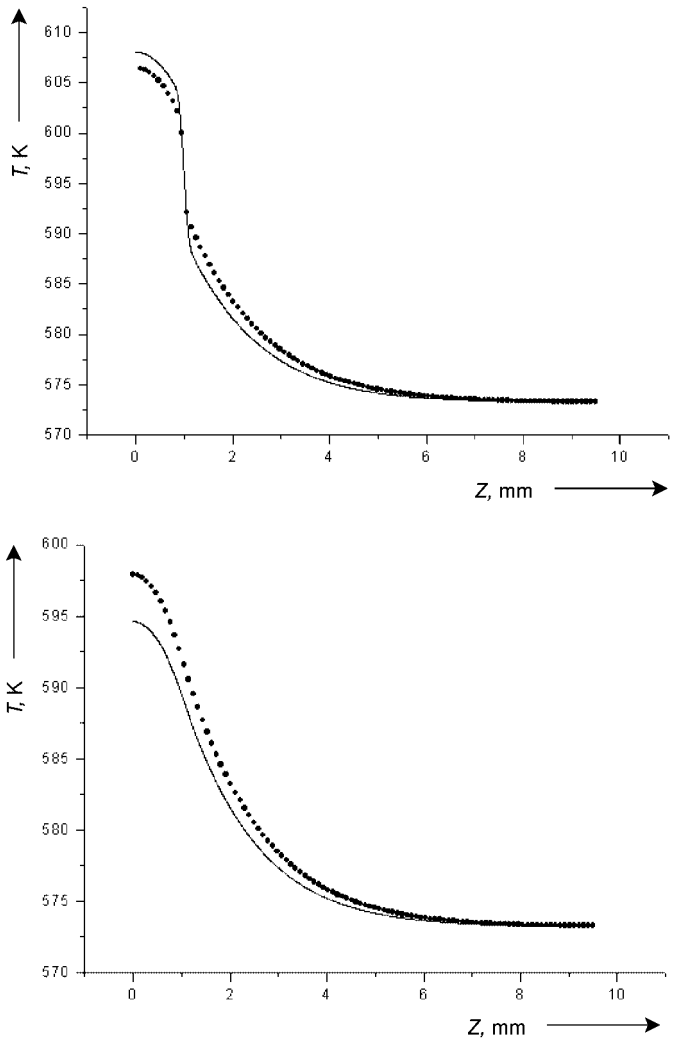

Figure 8. Fluid (left) and solid (right) temperatures in a multichannel stack as obtained from the full (dots) and the mean-field model (lines)

heat exchangers usually account only for transverse heat transfer from one fluid to another.

The performance of countercurrent micro heat exchangers was investigated, especially with respect to wall conduction effects and the choice of wall material [47], for a flow of nitrogen in channels with a length of 10 mm, a depth of 50 μm, and separating walls with a thickness between 125 and 500 μm. To reduce the computational effort, the heat exchanger geometry was only discretized in the axial direction, i.e., in the direction of the flow. In the direction perpendicular to the flow a fixed heat-transfer coefficient was used to describe the exchange of heat between the fluid and the channel walls.

When the thermal conductivity of the wall material is varied, characteristic temperature profiles for the channels with the hot and cold fluids and the channel wall are obtained. For very small values of the thermal conductivity, no heat is exchanged and the temperatures stay approximately constant. At intermediate wall thermal conductivities, almost linear temperature profiles are obtained, while at very high thermal conductivities the wall assumes a constant temperature and the temperatures of the fluids change rapidly in the entrance regions of

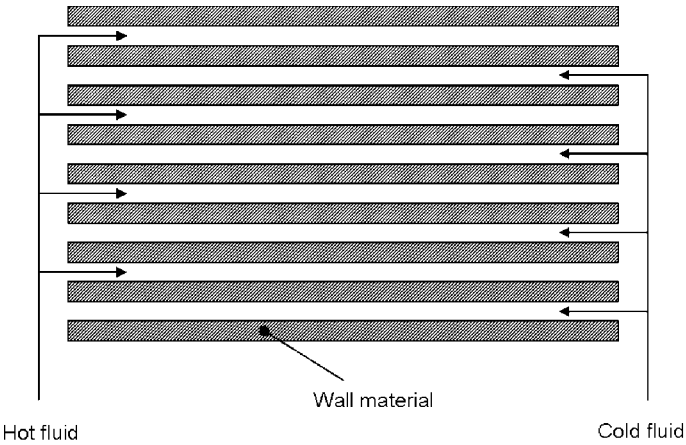

Figure 9. Schematic drawing of a countercurrent micro heat exchanger

the channels. A quantity characterizing the performance of a heat exchanger is the efficiency, which is the ratio of the transmitted heat and the maximum transmittable heat, given as

$$\varepsilon = \frac{T_{\text{in}}^{\text{hot}} - T_{out}^{hot}}{T_{\text{in}}^{\text{hot}} - T_{\text{in}}^{\text{cold}}} \tag{18}$$

in the case of equal volume flows of two identical incompressible fluids; T_{in} and T_{out} denote the temperatures at the channel in- and outlet, respectively. The heat-exchanger efficiency was calculated for a couple of characteristic geometries as a function of the wall thermal conductivity [47]. A typical result for equal volume flows is displayed in Figure 10. At low conductivities almost no heat is transferred, as expected. At high conductivities the wall assumes a uniform temperature due to axial heat conduction and the device becomes indistinguishable from a cocurrent heat exchanger, which is limited to an efficiency of 50 % at equal volume flows. At values around 1 W m^{-1} K^{-1} the curve assumes a maximum, as axial heat conduction inside the walls is suppressed to a sufficient degree and transverse conduction to the neighboring channels is still efficient. Hence, in typical micro heat exchangers the use of low-conductivity materials such as glass or polymers is preferable, and the common stainless steel materials are expected to reduce the efficiency.

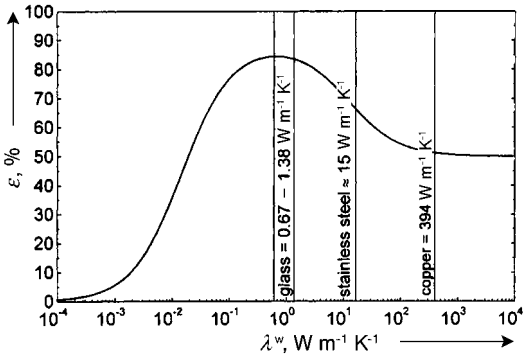

Figure 10. Heat-exchanger efficiency as a function of the thermal conductivity of the wall, taken from [47]

Heat losses to the exterior by conduction through the walls belong to the aspects not considered in the analysis in [47]. The vertical edges of the rectangular walls shown in Figure 9 are in thermal contact with the housing or the flow distribution manifold of the heat exchanger, and by these means heat is transferred to the surroundings. This effect was studied [48] by using a similar pseudo-1D model as [47], discretized only in the flow direction. Equal mass flow rates and identical incompressible fluids were assumed. The results were expressed as functions of the two independent dimensionless parameters of the problem, the conduction parameter

$$\lambda = \frac{kA}{\dot{m}\, c_p L} \tag{19}$$

and the number of transfer units

$$\text{NTU} = \frac{\alpha A}{\dot{m}\, c_p} \tag{20}$$

where k is the thermal conductivity of the wall material, A the cross-sectional area of a wall, $\dot{m}$ the mass flow, c_p the specific heat of the fluid, L the channel length, and α the heat-transfer coefficient between the fluid and the wall. As a result of the computational model, the heat-exchanger efficiency (or effectiveness) and the ratio of the conduction to the flow losses was obtained (Fig. 11). The conduction losses represent the heat flux dissipated to the surroundings via conduction through the walls, while the flow losses account for the fact that the cold fluid is not heated to the inlet temperature of the hot fluid (the thermodynamic limit) and vice versa. Figure 11 shows that the conduction losses are usually larger than the flow losses. Especially for a highly conductive wall material they exceed the flow losses by one order of magnitude or more. Thus, in combination with the results of Figure 10 the need for a low-thermal-conductivity construction material for countercurrent micro heat exchangers is underpinned in a most distinct way.

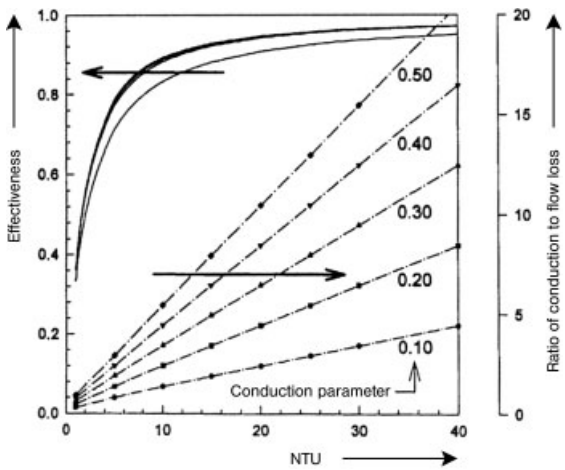

Figure 11. Effectiveness and ratio of conduction to flow loss as a function of NTU and conduction parameter for a countercurrent micro heat exchanger, taken from [48]

The studies of [47, 48] are both based on a description which assumes a vanishing transverse heat-transfer resistance within the channel walls, i.e., temperature gradients within the walls pointing in streamwise direction. A more elaborate model allowing for computation of the complete 2D temperature field in the wall material based on a Fourier expansion [49] confirmed that, with the exception of the close vicinity of the channel inlets and within the parameter range studied, transverse gradients are negligible, and thus corroborates the usefulness of pseudo-1D models for micro heat exchangers.

4. Mass Transfer and Mixing

Similar to heat transfer, fast mass transfer is one of the key aspects of microreactors. Again, due to the short diffusion paths, microreactors enable rapid mass transfer and a uniform solute concentration within the flow domain. Good control of reactant concentration throughout the whole reactor volume is a prerequisite for highly selective chemical reactions and helps to avoid hazardous operation regimes. In addition, overcoming mass-transfer limitations by rapid mixing allows the rapid intrinsic kinetics of chemical reactions to be exploited and enables higher yield and conversion. However, when dealing with liquid-phase reactions, fast mixing remains a challenge even at length scales of 100 µm or less due to the small diffusion constants in liquids.

Micromixing has been a very active field of research [50 – 52]. Not all of the micromixers reported in the literature are suitable candidates for microprocess engineering. On the one hand, the Reynolds numbers in microreactors are usually higher than those in Lab-on-a-chip devices which are often of the order of unity. On the other hand, microreactor processes are typically continuous flow processes for which batch-type micromixers are of little value. Owing to these differentiations, this section will mainly concentrate on mixing concepts and designs suitable for microprocess engineering and will highlight some specific simulation methods for micromixers.

4.1. Simple Mixing Channels

The simplest micromixer is the so-called mixing tee (Fig. 12, left). Two inlet channels merge into a common mixing channel where mixing of the two co-flowing fluid streams occurs. The mixing characteristics of T-type mixers were investigated by using CFD methods [53]. Mixing of gases in a channel of 500 µm width was considered and certain geometric parameters such

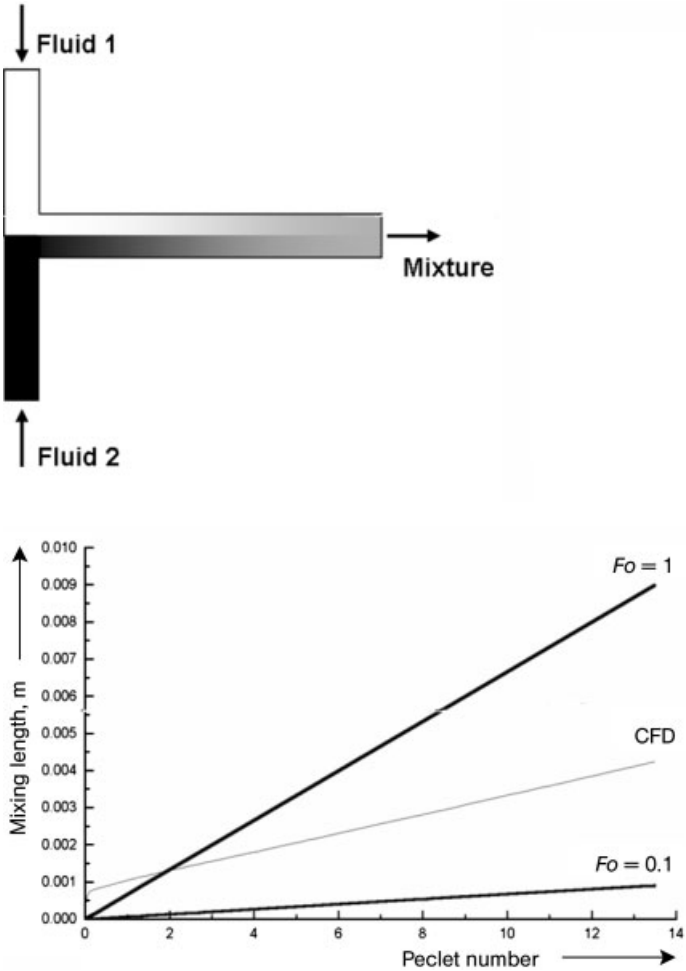

Figure 12. Schematic design of a mixing tee (left) and CFD results for mixing of gases in a channel of 500 µm width and 300 µm depth (right), taken from [53]

as the aspect ratio of the mixing channel or the angle at which the two inlet channels meet were varied. To quantify mixing, a mixing length was defined as the length in flow direction after which the gas composition over all positions of a channel cross section deviates by no more than 1 % of the equilibrium composition. The CFD results for the mixing length as a function of Péclet number $Pe = \bar{u}\ d_{\mathrm{h}}/D$ are displayed on the right side of Figure 12, where $\bar{u}$, d_{h}, D are the average flow velocity, hydraulic channel diameter, and molecular diffusivity, respectively. The maximum Reynolds number considered was about 17.

In addition to the CFD results, estimates of the mixing length based on the Fourier number

$$\mathrm{Fo} = \frac{Dt}{l^2} \tag{21}$$

are shown in the figure. The Fourier number relates a residence time t in the mixing channel to the molecular diffusion constant D and a characteristic length scale l, which is the width of the channel. For a given value of Fo and a given flow rate the length along the mixing channel necessary to achieve the corresponding Fourier number was determined. As is apparent from Figure 12, the Fourier number is a reasonable indicator for mixing, which occurs for Fourier numbers between 0.1 and 1.0. Consequently, l^2/D is an order-of-magnitude representation of the mixing timescale. The linear increase of the CFD-based

mixing length as a function of Péclet number points to a very simple mixing mechanism by diffusion between co-flowing fluid lamellae. Obviously, complex convection-dominated mixing mechanisms (e.g. driven by swirls or recirculating flows) are absent in the simple mixing-tee configuration for the range of Reynolds numbers studied.

When the Reynolds number is increased, a regime with a flow pattern of broken symmetry is reached [54]. At a Reynolds number of about 140 the symmetric flow pattern becomes unstable and an engulfment flow with an increased area of material interface between the two fluids is found. This is illustrated in Figure 13, which shows maps of the computed concentration field over cross sections of the mixing channel for Reynolds numbers ranging from 119 (a) to 239 (f). The increase in material interface is related to an increase in mixing efficiency, and thus a mixing tee is more efficient than predicted by simple models accounting for mass transfer through an interface located in the symmetry plane of the mixer.

In cases where symmetry breaking and the corresponding engulfment flows are absent (i.e. at sufficiently small Reynolds numbers) the steady-state concentration field in a mixing tee can be computed based on an analytical approach. The underlying assumption is that the flow moves with uniform velocity. Ref. [51] gives the following Fourier series expansion for the concentration field in a mixing channel of width w with two streams of equal viscosity and a volume flow ratio of $r/(1-r)$

$$\frac{c(x,y)}{c_0} = r + \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{\sin nr\pi}{n}\cos\frac{n\pi y}{w} \exp\left(-\frac{2n^2\pi^2}{Pe+\sqrt{Pe^2+4n^2\pi^2}}\frac{x}{w}\right) \quad (22)$$

where the concentrations in the two streams at the channel entrance are $c = c_0$ and $c = 0$ and Pe is the Péclet number based on the channel width. The coordinates in length and in width direction of the microchannel are denoted as x and y, respectively. With such an analytical expression the evaluation of mixing performance in a simple microchannel is very time efficient. However, an even simpler analysis based on the Fourier number already gives the correct order of magnitude for the mixing length, as indicated in Figure 12.

4.2. Chaotic Micromixers

The performance of mixing tees is usually not sufficient for liquid-mixing applications, since corresponding diffusion constants are of the order of 10^{-9} m^2/s or smaller, and mixing times far below one second may be required. As described above, mixing in straight channels occurs mainly by diffusion between co-flowing streams with a characteristic time constant of l^2/D. A reduction of the mixing timescale can be achieved if diffusion is superposed by convective transport. On the macroscale, convective transport is often due to turbulent eddies, and only in the very final stages of the mixing process does diffusion play a role. Unfortunately, in microscopic geometries flow is usually laminar, and a different convection mechanism must be devised to speed up mixing.

Much attention has been devoted to creating flow schemes in micromixers which enhance the mixing performance by convective transport. In that context it was found that chaotic flows bear a great potential for speeding up mixing. The corresponding principle of chaotic advection has been studied for quite some time [55], but it has only more recently been implemented in microchannel flows. Chaotic advection is synonymous with complex flow patterns in which extreme stretching and backfolding of fluid volumes occurs. It is very difficult to numerically study chaotic micromixing of liquids. The reason lies in the discretization artifacts from which most CFD schemes suffer, specifically the phenomenon of numerical diffusion [56, 57]. Owing to discretization errors artificial fluxes are induced which, for specific differencing schemes, take the form of diffusive fluxes and thus artificially increase the diffusion constant. The magnitude of numerical diffusion depends on the cell-based Péclet number (evaluated with the extension of a grid cell as length scale) and the relative orientation of the grid cells and the flow velocity. Specifically, if the flow is parallel to the grid lines, the diffusivity perpendicular to the flow direction is zero. In the case of a simple, regular flow pattern it is often possible to align the computational grid with the velocity field and thereby suppress numerical diffusion without having to prohibitively increase grid resolution. However, in the case of a chaotic flow

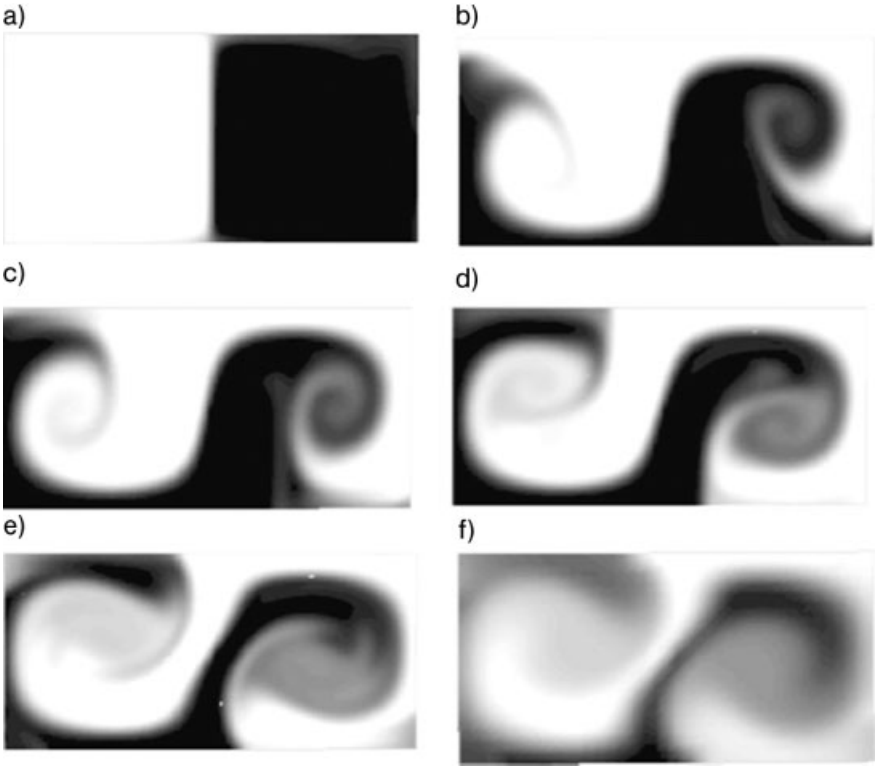

Figure 13. Concentration field over the channel cross section of a mixing tee for Reynolds numbers of 119 (a), 139 (b), 146 (c), 159 (d), 186 (e), and 239 (f), reproduced from [53]

this is no longer possible and numerical diffusion often severely falsifies the computational results.

To evaluate the performance of chaotic micromixers, the standard CFD methods for solving convection – diffusion equations must be abandoned and replaced by a method which is largely free of discretization artifacts. Such a method is Lagrangian particle tracking, which is based on tracking the paths of massless particles in a given flow field. Assuming that the flow velocity u_i is given, the particle trajectories are obtained as solution of the equation

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = u_i \tag{23}$$

where x_i denotes the position vector of a particle. The mixing process is then studied by seeding the inlet of the mixer with a multitude of particles and recording how these particles have been redistributed by the flow at a downstream position. The fact that Equation 23 does not contain any contribution of Brownian particle motion is often not a real drawback. If the flow itself displays a large degree of chaoticity, much of the redistribution of the initial concentration field is done solely by convective transport, and diffusion only completes the mixing process. In a coarse-grained description in which concentrations are evaluated by computing the particle numbers in specific grid cells, the inclusion of Brownian motion would not modify the results significantly in many cases.

The computational effort for Lagrangian particle tracking can be reduced if the mixing section exhibits some periodicity. The structures inducing the chaotic flow are often arranged in a certain spatial period for which the flow field may be computed using periodic boundary conditions. Particles are then seeded onto the entrance plane of this section of the mixing channel and tracked up to its exit plane. The fluid streamlines map each point of a channel cross section (x_k, y_k) to another point (x_{k+1}, y_{k+1}) of the corresponding cross section after one period of the fluidic structures, as shown in Figure 14.

Such a so-called Poincaré map can be written formally as

$$(x_{k+1}, y_{k+1}) = P\left[(x_k, y_k)\right]. \tag{24}$$

The function P can be computed either from an analytical or a numerical representation of the flow field, making use of Equation 23. In such a way, a 3D convection problem is essentially reduced to a mapping between two-dimensional Poincaré sections. The complete mixing process can be analyzed by successive application of the Poincaré map.

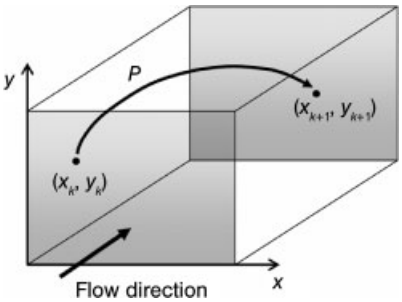

Figure 14. Section of a mixing channel with a map P connecting the points of two cross-sectional planes

Lagrangian particle tracking techniques have been used to evaluate the performance of several chaotic micromixers. One of the first micromixers for which convincing evidence for chaotic mixing was given is based on a design known as staggered herringbone mixer (SHM) [58]. The corresponding mixer geometry is shown in Figure 15. The bottom of the channel of height h and width w contains a staggered arrangement of grooves, whereby a fraction p of the bottom channel wall has grooves inducing a helical flow with a left-handed recirculation, and the remaining bottom wall fraction induces a right-handed recirculation. Schematic views of the channel cross section including projections of the streamlines are shown at the top of Figure 15. If $p \neq 1/2$, the vortex pattern is asymmetric, and a superposition of two patterns exhibiting the larger of the two vortices on the left and on the right side, respectively, could result in chaotic flow. In a generalized sense, this is an implementation of the "blinking-vortex principle" [59]: two vortex structures can be superposed in an alternating fashion to create a chaotic flow pattern. In the SHM the superposition is achieved by the alternating, staggered groove patterns shown in Figure 15. The lower part of the figure shows channel cross sections with two streams of fluorescent and clear solutions after 0, 0.5, and 1 cycle. The images were recorded using a confocal microscope. The two different vortices are clearly visible, and the third frame shows first indications of a chaotic disturbance of the flow.

Mixing in the SHM has been analyzed using Lagrangian particle tracking [60]. By comparison with a mixer containing only simple groove structures (as opposed to staggered herringbone structures) evidence was given for the superior performance of chaotic micromixers. The Poincaré sections obtained from particle tracking were compared to the confocal micrographs of Figure 15 [58]. The computed particle distributions (Figure 16) show strong similarities to the distributions of fluorescence intensity, and thus prove that mixing in complex flows can realistically be described by Lagrangian particle tracking. With standard CFD techniques resting on the solution of a convection – diffusion equation for the concentration field this is often not possible, since discretization errors would have blurred the thin lamellae visible in Figure 16.

Lagrangian particle tracking has also been applied to analyze chaotic mixing in a number of other channel designs. Simulation Studies [61] on a twisted microchannel [62] showed that for Reynolds numbers above 10 mixing is mostly chaotic, apart from a few unmixed, regular islands. Cross-channel micromixers have been studied by a number of authors [63 – 65]. In such devices multiple side channels intersect with the main mixing channel. An oscillating pressure is imposed on the side channels such that the time-averaged flow rate entering the main channel is zero. The flow entering from the side channels distorts the material interface of the two fluids to be mixed, thereby increasing the interfacial area and promoting mixing. The analysis of mass transfer based on Lagrangian particle tracking has shown that, for suitable values of the pressure excitation amplitude and frequency, chaotic mixing is induced [63].

Many of the studies on micromixing reported in this section are more or less geared towards Lab-on-a-chip applications, i.e., the target Reynolds number is about 10 or smaller. In microreactors throughput is often a more impor-

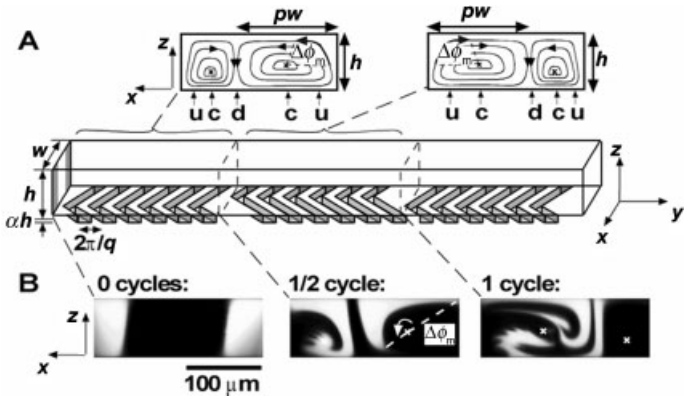

Figure 15. Micromixer geometry with staggered groove structures on the bottom wall, as considered in [58] (reprinted with permission, © 2002, AAAS). The top of the figure shows a schematic view of the channel cross section with the vortices induced by the grooves. At the bottom, confocal micrographs show the distribution of two liquids over the cross section. Flow is from left to right.

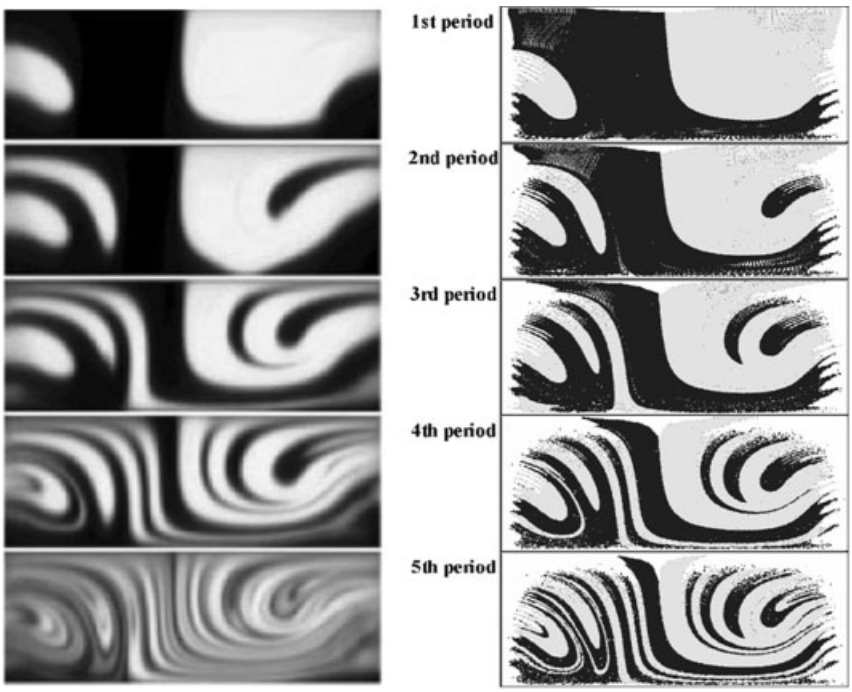

Figure 16. Comparison of confocal micrographs [58] (left) with the Poincaré sections obtained from Lagrangian particle tracking [60] (right).

tant issue than in Lab chips and, correspondingly, Reynolds numbers are higher and may easily reach values of several hundreds. A quite simple micromixer which is well suited for this Reynolds number regime was analyzed theoretically and experimentally [66]. The design of the mixer is shown at the bottom of Figure 17. The mixing channel consists of a number of arclike segments of alternating curvature, two of which are shown in the figure. As discussed in Section 2.2, a helical flow develops in such curved channels, with vortices appearing in the projection of the velocity field onto the channel cross section. For Dean numbers larger than 140 the flow exhibits four counterrotating vortices, as indicated by the CFD results shown in the middle of Fig-

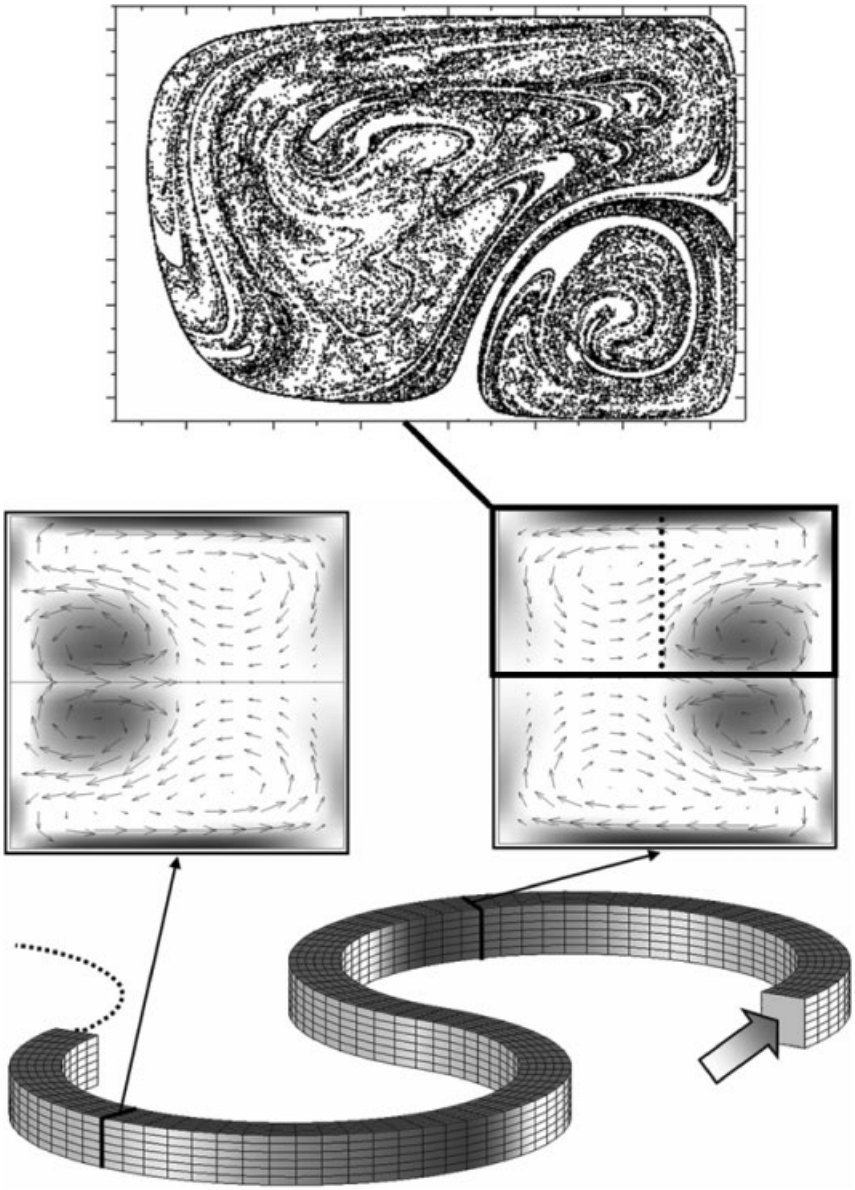

Figure 17. Two segments of a meandering mixing channel [66] (bottom) and CFD results showing the projected velocity fields on different cross-sectional planes (middle). The upper part of the figure shows the positions of massless particles initially seeded along the material interface after six channel segments.

ure 17. In the mixer proposed by [66] a chaotic flow is created by alternating between the flow patterns in different channel segments, with the smaller pair of vortices either appearing in the right or the left half of the cross-sectional plane.

To study the mixing performance, a modified version of the usual Lagrangian particle tracking was developed. Usually, two types of tracer particles are distributed over the entire entrance plane and the mixing quality is computed by counting the number of particles in grid cells on a plane located at a downstream position. Based on this method it is difficult to obtain reliable results for states close to complete mixing. In an alternative method [66] only the material interface is seeded with particles (indicated as the

dotted line in the upper half of one channel cross section in Fig. 17) and the length of the intersection of this interface with cross-sectional planes at various downstream positions is computed. This length scale should provide a good measure for mixing, since all the mass transfer proceeds through the material interface, and can be computed with high accuracy, since the same number of particles usually distributed over a plane can now be distributed along a line.

The upper part of Figure 17 shows the distribution of particles along the material interface after six channel segments for a Dean number of 200. The particles are distributed over almost the entire cross-sectional plane and thus document the pronounced and very complex stretching and folding of the interface. For Dean numbers above 100 an exponential growth of interfacial length scale as a function of residence time in the channel could clearly be proven. From these results it is apparent that mixing proceeds in a chaotic manner. This statement was underpinned by experimental results which showed that the position of close-to-complete mixing along the channel is virtually independent of the flow rate.

4.3. Multilamination Micromixers

The chaotic micromixers described in Section 4.1 offer a quite elegant solution to the micromixing problem, as complicated manifolds for distributing the fluid streams are no longer required and mixing is achieved by the intrinsic structure of the flow field. However, if the flow is chaotic in one part of the channel domain and regular in the other, unmixed islands, which usually deteriorate process performance, will remain. Thus, before using a chaotic micromixer in practice its mixing performance should be carefully evaluated over the whole range of Reynolds numbers considered. In comparison, multilamination mixers often require much more complex distribution manifolds and channel structures, but their mixing performance can usually be better controlled and predicted over the operation range envisioned.

The principle of mixing by multilamination relies on the creation of a multitude of thin fluid lamellae between which mixing occurs by diffusion. Thus, the mixing times achievable with multilamination mixers are directly related to the diffusion timescales between neighboring fluid lamellae. Figure 18 exemplifies the basic flow structure in a multilamination mixer: A multitude of fluid streams is guided through a flow chamber of varying width. For the mixer in the figure the streams are focused into a narrow channel by which the characteristic diffusion path is drastically reduced and mixing is enhanced.

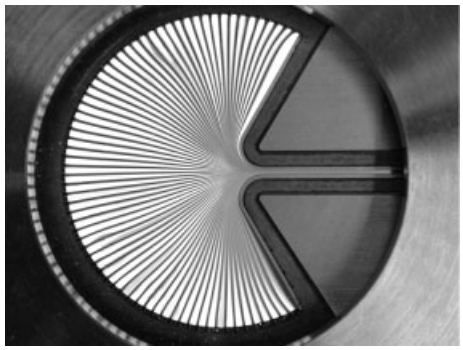

Figure 18. Multilamination mixer focusing a multitude of fluid streams through an inlet chamber into a narrow mixing channel. Alternating fluid inlets are distributed over a circular arc-shaped manifold.

In a similar way as in Section 4.1 an analytical formula can be derived which allows for a fast computation of the concentration field in a multilamination mixer. By transforming the spatial coordinate in streamwise direction (x in Eq. 22) into a time coordinate by dividing the spatial differerential by the average flow velocity and integrating from the inlet to the outlet of the mixer, a 2D spatial problem is reformulated as a problem with one space and one time coordinate. Thus, in a frame of reference co-moving with the flow and assuming a plug-flow profile, the convection – diffusion problem is transformed into a pure diffusion problem. The resulting diffusion equation can be solved by the usual Fourier-transform techniques. A corresponding analytical model for multilamination mixing has been developed [67] which takes into account changes of the channel width along the flow path, i.e., it incorporates effects of focusing or defocusing, with the former often being used to speed up mixing. Specifically, a coordinate s orthogonal to the fluid lamellae is introduced which may be a curvilinear coordinate in regions where focusing or defocusing occurs. It is assumed that s

lies between $-w/2$ and $w/2$, where w is a measure for the channel width. The concentration field is then expanded in harmonics

$$\frac{c(s,t)}{c_0} = \frac{1}{2} + \sum_{n=0}^{\infty} a_n(t) \sin\frac{k_n s}{w}, \quad k_n = (2n+1)\pi \tag{25}$$

where only the sine terms contribute if the total number of fluid lamellae is even and initial concentrations of c_0 and 0 are considered. The time dependence of the concentration field arises from the fact that a frame of reference co-moving with the flow is chosen, as described above. Inserting this expansion into the diffusion equation in the co-moving frame of reference allows $a_n(t)$ to be determined as

$$a_n(t) = a_n(0) \exp\left(-Dk_n^2 \int_0^t \frac{\mathrm{d}t'}{w^2(t')}\right), \tag{26}$$

where D is the molecular diffusivity. The function $w(t)$ incorporates the changes in channel width. In contrast to a channel of constant width an integral must be evaluated to determine the Fourier coefficients. This integral represents the history of the mixing process and accounts for the changes in mixing speed due to width variations.

Clearly, the model of Equations 25 and 26 should also be applicable to a simple mixing channel with only two fluid streams as considered in Section 4.1, i.e., Equation 22 should be recovered when transforming to a formulation with two spatial coordinates instead of a space – time description. However, the difference between the two formulations is that Equation 22 also takes into account diffusion in streamwise direction, an effect that becomes important at small Péclet numbers. Nevertheless the model for multilamination mixing is useful in most practical cases, since diffusion in streamwise direction is usually negligible. This model is free of numerical diffusion and thus not only allows faster, but also more reliable prediction of mixing performance in specific cases. The fast evaluation of mixing performance allows parameter studies to conducted which would have been extremely time consuming when using standard CFD techniques [68].

Reducing the solution of the convection – diffusion equation for the concentration field to a 1D problem in the co-moving frame-of-reference clearly reaches its limits if the fluid lamellae become deformed and are no longer arranged in parallel. Such a situation may occur if the channel depth is not small compared to its width and the design comprises a sudden expansion or contraction of the flow passage. Multilamination mixers with cornerflow geometries have been analyzed by CFD techniques, and situations where the multilamination pattern suffers strong deformations identified [69]. The complex flow patterns found in such mixers may also mislead experimentalists who attempt to evaluate the mixing performance by photometric techniques based on the diffusion of a solute [69].

Creating a multitude of lamellae such as in Figure 18 usually requires a quite complex flow distribution manifold. There is, however, an alternative multilamination principle which enables multiplication of the number of lamellae along the mixing channel: split-and-recombine (SAR) mixing. An SAR mixer consists of successive mixing elements each of which causes a doubling of the number of lamellae. An example of such a mixing element is shown in Figure 19. The flow entering the element in positive z-direction is split into two substreams which are then recombined. Thereby the number of fluid lamellae is increased from two to four. For n successive mixing elements, the number of lamellae is increased to 2^{n+1} and the diffusion paths are reduced correspondingly. The exponential reduction of diffusion paths resembles mass transfer in chaotic flows, where fluid volumes are stretched and folded such that the width of fluid filaments decreases exponentially. For this reason SAR mixers are usually very efficient, without the need for incorporating very small channels. Furthermore, in contrast to some chaotic mixers SAR mixers guarantee a largely uniform mixing quality over the entire channel cross section.

For the channel design shown in Figure 19 it could be shown that lamellae multiplication works in a close-to-ideal manner for Reynolds numbers below about 30 [70]. For higher Reynolds numbers inertial forces become important and cause deformation and finally merging of lamellae of the same type. Altogether, experiments indicate that at comparatively small Reynolds numbers the design shown enables an efficient liquid mixing at low pressure drops, and thus the SAR mixer is especially qual-

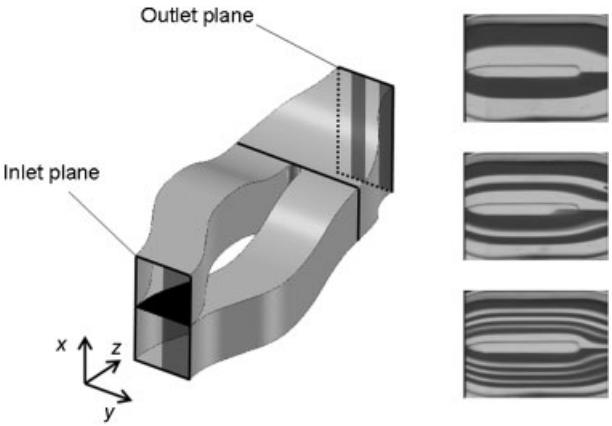

Figure 19. Mixing element of an SAR mixer (left) and optical micrographs showing lamellae multiplication in the first three mixing elements (right)

ified for the processing of highly viscous liquids [70].

As far as the computation of concentration fields in SAR mixers is concerned, the challenges are greater than in ordinary multilamination mixers. The reason is that the number of lamellae does not remain constant; rather, there are discontinuous transitions between patterns with n and $2n$ lamellae. A strategy to deal with this problem has been formulated [67]: based on solving a 1D diffusion equation in a co-moving frame-of-reference (Eqs. 25 and 26), the mixing element shown in Figure 19 is subdivided in two sections. In the first section, the flow passes through two separate branches, and in the second section it only passes through one branch. In each of these sections the time evolution of the Fourier coefficients is computed by Equation 26. At the interface, the downstream Fourier coefficients are obtained from a sum involving the upstream Fourier coefficients. In addition, the evolution of the concentration field in a complete SAR mixer comprising several successive mixing elements is obtained in an iterative manner from the analysis of a single mixing element, i.e., the Fourier coefficients on the outlet plane shown in Figure 19 are identified with the coefficients on the inlet plane of the next element. Thus the degree of mixing was computed as a function of position along the channel, and reasonable agreement with experimental data was found [67]. Thus, the usefulness of the 1D model for multilamination mixing is corroborated by experimental data, and reasonable results can be obtained even for such complex situations involving flow passages of varying width as well as flow splitting and recombination.

There are alternative flow topologies for SAR mixers which have been proven to cause a multiplication of fluid lamellae. The design of mixing element shown in Figure 19 was especially optimized with regard to extending the Reynolds number range in which the SAR principle is implemented, consistent with the throughput requirements typical for microreactors. The disadvantage is that this mixing element poses major challenges for microfabrication. When only very small Reynolds numbers must be considered the design constraints are less severe, and a design which is easier to fabricate has been proposed [71]. The flow topology is such that a doubling of the number of fluid lamellae is achieved after two SAR steps, and the functionality with respect to lamellae multiplication has been proven by using dyed aqueous solutions. In the same spirit as above, the model of Equation 25 and 26 can be used to compute the evolution of concentration fields in such a mixer if the upstream and downstream Fourier coefficients at the branching points are matched with each other.

4.4. Hydrodynamic Dispersion

Mass transfer of a solute dissolved in a fluid is not only the fundamental mechanism of mixing processes, it also determines the residence-time

distribution in microfluidic systems. As mentioned in Section 2.3, in many applications it is desirable to have a narrow residence-time distribution of concentration tracers being transported through a microfluidic system. An initially narrow concentration of tracer will suffer broadening (i.e., dispersion) due to two different effects. First, in some regions of the flow domain of a system the fluid velocity will be smaller than in others, and this leads to a longer residence time of molecules being transported preferably through these parts of the domain. However, due to Brownian motion the molecules will also sample some of the other regions with higher flow velocity. Hence, molecular diffusion may reduce the dispersion of a concentration tracer. On the other hand, by diffusion an initially localized concentration tracer in a fluid at rest will be dispersed. From these arguments it becomes clear that hydrodynamic dispersion depends on a quite subtle interplay of convective and diffusive mass transfer, and the evolution of a concentration tracer as it is transported through the flow domain depends on various factors such as the flow profile, the magnitudes of the flow velocity, and the diffusion constant.

The key analysis of hydrodynamic dispersion of a solute flowing through a tube [72, 73] assumed a Poiseuille flow profile in a tube of circular cross section and showed that for large enough times the dispersion of a solute is governed by a one-dimensional convection – diffusion equation

$$\frac{\partial \bar{c}}{\partial t}+\bar{u}\,\frac{\partial \bar{c}}{\partial x}=D_e\frac{\partial^2 \bar{c}}{\partial x^2} \tag{27}$$

where $\bar{c}$ is the concentration averaged over the cross section of the tube, $\bar{u}$ the average velocity, and D_e an effective diffusivity, also denoted dispersion coefficient, which is given by

$$D_e=D+\frac{\bar{u}^2R^2}{48D} \tag{28}$$

where D is the molecular diffusivity, and R the radius of the tube. The factor 1/48 multiplying the velocity-dependent term is generic for tubes of circular cross section and is modified when other geometries are considered. In many cases the second term, which can be rewritten as $DPe^2/48$, dominates over the first, which is a purely diffusive contribution. Hence, due to convection a concentration tracer is usually dispersed much more strongly than it would have been by diffusion alone. A notable feature of Equation 27 and 28 is their independence of any initial condition. Independent of how the solute is distributed over the channel cross section and along the channel initially, the Taylor – Aris [72, 73] description is valid in the limit of long times ($t\cdot\cdot$). When exactly this limit is reached with a given accuracy depends on the initialization of the concentration field. A rough guideline is provided by the Fourier number of Equation 21, evaluated with the tube radius as length scale. The Fourier number can be regarded as a dimensionless time coordinate which compares the actual time to the time a molecule needs to sample the cross-sectional area of the tube. The validity of the description should be related to the condition that the Fourier number assumes values on the order of 1 or larger.

The above analysis was extended to arbitrary time values for the dispersion of an initially plug-like profile [74], i.e.

$$c\,(x,r,0)=\{\begin{matrix} c_0\ (|x|\leq l/2) \\ 0\ (|x|>l/2)\end{matrix} \tag{29}$$

where r is the radial coordinate of the tubular geometry, and l the length of the plug. A generalized evolution equation was derived for the area-averaged concentration of the form

$$\frac{\partial\,\bar{c}}{\partial t}=\sum_{n=1}^{\infty}k_n\,(t)\,\frac{\partial^n\,\bar{c}}{\partial x^n} \tag{30}$$

which is valid without any restriction on t. The derivatives in the infinite series appearing on the right-hand side are multiplied by time-dependent dispersion coefficients k_n. In the Taylor – Aris limit, all of the dispersion coefficients except k_1, which describes the convection of the tracer with the flow, and k_2, which determines the spreading of the tracer, are negligible. When moving to smaller times, the time-dependence of k_2 needs to be taken into account, while all higher dispersion coefficients are still negligible [74]. Only at very small times do the higher dispersion coefficients become important. For the case considered, k_2 can be regarded as time-independent for Fourier numbers greater than about 0.5 [74].

Even if the results discussed above highlight some of the most important aspects of hydrodynamic dispersion, they were based on cylindrical ducts, which are not the generic geometry

used in the field of microfluidics. In chemical microprocess technology, tubular sections are used to connect different units, but the channels contained in microreactors typically have a rectangular or tublike cross section. Dispersion in rectangular channels was studied in detail [75]. The evolution equation (Eq. 27) is still valid in this case, but the expression for the dispersion coefficient (Eq. 28) needs to be modified. While a simple analytical expression for the dispersion coefficient related to flow between parallel plates could still be obtained [76], the corresponding expression for rectangular channels is a complicated series expansion. This is not surprising, since the exact form of the flow profile in a rectangular channel is given in form of an infinite series as well. A simpler relationship approximates the exact expression for the dispersion coefficient in rectangular channels within an error of 10 % [77]. In addition, the tub-like channel cross sections which are typically obtained by isotropic etching processes were considered by employing a numerical scheme that allowed computation of the dispersion coefficient. On this basis, different channel geometries were compared, and favorable and less favorable designs identified.

In cases where hydrodynamic dispersion and the corresponding broadening of residence-time distributions deteriorate the performance of a process, the question arises which channel design minimizes dispersion. Already from the Taylor – Aris analysis it becomes clear that enhanced mass transfer perpendicular to the main flow direction reduces the broadening of concentration tracers. Such a mass transfer enhancement can be achieved by the secondary flow occurring in a curved channel. This aspect was investigated for ducts of circular cross section [78] under the assumption that the diameter of the duct is small compared to the radius of curvature, and the convection – diffusion equation for the concentration field was solved numerically. More specifically, a two-dimensional problem defined on the cross-sectional plane of the duct was solved based on a combination of a Fourier series expansion and an expansion in Chebyshev polynomials. The solution is of the general form

$$D_e^{cur} = D\left[1 + Pe^2 f(K, Sc)\right], \tag{31}$$

where D_e^{cur} is the dispersion coefficient in curved ducts, D the molecular diffusivity, Pe the Péclet number of the flow, and the function f depends on the Dean number K defined in Equation 7 and on the Schmidt number Sc, which is the ratio of the kinematic viscosity and the diffusivity. The following asymptotic behavior for the ratio of the dispersion coefficients in curved and straight ducts was found [78]:

$$\frac{D_e^{cur}}{D_e}\begin{cases} = 1 & K \to 0 \\ \propto K^{-1} & K \to \infty \end{cases}. \tag{32}$$

In curved channels secondary flow patterns of two or four counterrotating vortices are formed. These vortices redistribute fluid volumes in a plane perpendicular to the main flow direction. Such a transverse mass transfer reduces dispersion, a fact reflected in the K^{-1} dependence in Equation 32 at large Dean numbers. For small Dean numbers, the secondary flow is negligible, and the dispersion in curved ducts equals the Taylor – Aris dispersion of straight ducts.

Similar to the case of mixing, the transverse redistribution of matter in a curved channel can be improved in a sequence of channel segments of alternating curvature. By superposition of different vortex patterns a chaotic flow is induced [66]. By Lagrangian particle tracking in combination with a CFD computation of the flow field it can be shown that the residence-time distribution of tracers in the channel built from segments as shown in Figure 17 becomes very narrow and approaches a deltalike distribution of zero width. In practice, however, the width of the residence-time distribution is limited by the molecular diffusivity of the tracers, which was neglected in the simulations [66]. Nevertheless, a substantial reduction of hydrodynamic dispersion can be achieved in a meandering channel when compared to a straight channel.

5. Chemical Kinetics

Most plants or reactors in chemical microprocess technology inevitably contain a unit in which chemical conversion takes place. The goal might be to produce a fine chemical with a high yield and selectivity or to screen a large number of reactions in parallel. Hence, a thorough understanding of chemical kinetics is a key requirement for the successful design and optimization of micro reaction devices. For this purpose, reliable models of reaction kinetics cou-

pled to the transport equations of momentum, heat, and matter are needed.

The type of kinetic model to be used depends on the type of reaction considered. For a homogeneous reaction occurring in the bulk of the fluid, a power-law kinetic model is often appropriate (see, e.g., [79]). In such models the rate of a certain reaction depends on a product of powers of the species concentration. On the other hand, heterogeneously catalyzed reactions are often conducted in microreactors. In a strict sense, power-law kinetics do not capture the dynamics of such processes over the full range of pressure, temperature, and concentrations. Instead, a more complicated kinetic model of, e.g., the Langmuir – Hinshelwood type [80] would have to be used. Nevertheless, power-law kinetics is frequently applied to heterogeneously catalyzed processes in a limited parameter range to simplify the description.

Independent of the specific modeling strategy, the kinetic equations often exhibit a nonlinear dependence on the species concentrations and a reaction rate rapidly increasing with temperature. In combination with the transport equations for mass, momentum, and heat, the resulting numerical problem is usually challenging due to the nonlinearities and the multitude of timescales involved. For this reason, methods are needed to eliminate some of these difficulties and to simplify the numerical structure.

5.1. Numerical Methods for Reacting Flows

The solution of the species concentration equations in combination with the momentum and the enthalpy equation generally requires an iterative procedure. A rough sketch of the numerical structure of a stationary reacting-flow problem is given as

$$\begin{bmatrix} A_{cc} & A_{cu} & A_{cT} \\ A_{uc} & A_{uu} & A_{uT} \\ A_{Tc} & A_{Tu} & A_{TT} \end{bmatrix} \begin{bmatrix} c \\ u \\ T \end{bmatrix} = \begin{bmatrix} b_c \\ b_u \\ b_T \end{bmatrix}, \tag{33}$$

where c, u, and T denote the vectors of concentration, velocity, and temperature fields, respectively. Owing to the nonlinear nature of the problem, the coefficients of the different matrices $A_{\alpha\beta}$ still depend on the unknowns c, u, and T. The cross-coupling between different field quantities is provided by those matrices $A_{\alpha\beta}$ with $\alpha \neq \beta$. The set of nonlinear algebraic equations is solved iteratively, i.e., starting with an initial guess the approximation is successively improved until convergence is reached. Depending on the nature of the chemical reaction term entering the species-concentration equation, different strategies may be applied to solve Equation 33. For intrinsic kinetics characterized by a much shorter timescale than transport of momentum, heat, and matter, it is often preferable to set up an iteration scheme in which a number of iterations of the species-concentration equation are performed during one iteration cycle of the remaining equations. However, for a fast reaction which is limited by heat and mass transfer (e.g. in a situation where the reactants are not premixed), comparable iteration cycles of the species-concentration equation and the remaining equations might be sufficient.

Apart from the coupling of chemical kinetics to the transport equations, the chemical reaction dynamics themselves may pose numerical challenges when a number of different reactions are superposed. In such a case the rate of disappearance of a chemical species i can be written as

$$-\frac{\dot{n}_i}{V} = R_{ij} r_j, \tag{34}$$

where r_j is the rate of the jth reaction, and R_{ij} a matrix defining how a specific reaction contributes to a change in concentration of the chemical species involved. For brevity the Einstein convention of summation over repeated indices was used. Quite frequently it occurs that the timescales characterizing the different reactions vary by orders of magnitude, such that the fast reactions are already completed while the slow reactions have not yet progressed to any appreciable degree. The corresponding stiff differential equations are usually solved by an implicit time-integration scheme which allows comparatively large time steps without suffering from numerical instabilities or predicting unrealistic asymptotic states [81]. However, implicit time integrators involve the solution of a (generally nonlinear) algebraic system of equations for each time step, which is done by some iterative scheme such as Newton's method. For reaction systems with a broad spectrum of timescales these iteration schemes can fail to converge, with the consequence that very small time steps must be

chosen. Such a situation is related to high computational costs, and methods are needed to simulate extremely stiff reaction systems more efficiently.

Methods based on the partitioning of a reaction system into fast and slow components were proposed [82 – 84]. A key assumption made in this context is the separation of the space of concentration variables into two orthogonal subspaces Q_s and Q_f spanned by the slow and fast reactions. With this assumption the time variation of the species concentrations is given as

$$-\frac{\dot{n}_i}{V}=(Q_{s,j})_i\left(\dot{y}_s\right)_j+(Q_{f,j})_i\left(\dot{y}_f\right)_j. \tag{35}$$

The notation is such that $(Q_{\mathrm{s},j})_i$, $(Q_{\mathrm{f},j})_i$ denote the ith component of the jth basis vector in the subspace of slow and fast reactions, respectively. The corresponding expansion coefficients are $\left(\dot{y}_s\right)_j$ and $\left(\dot{y}_f\right)_j$, respectively, and are expressed by the reaction rates via

$$\left(\dot{y}_s\right)_i=\left(Q^T_{s,j}\right)_i R_{jk}r_k, \tag{36}$$

$$\left(\dot{y}_f\right)_i=\left(Q^T_{f,j}\right)_i R_{jk}r_k. \tag{37}$$

If the timescale of the fast reactions is much shorter than that of the slow reactions, it can be assumed that the former are completed at an initial stage of the latter. Mathematically, this assumption reads

$$\left(Q^T_{f,j}\right)_i R_{jk}r_k=0. \tag{38}$$

Equation 38 represents a set of algebraic constraints for the vector of species concentrations expressing the fact that the fast reactions are in equilibrium. The introduction of constraints reduces the number of degrees of freedom of the problem, which now exclusively lie in the subspace of slow reactions. In this a way the fast degrees of freedom have been eliminated, and the problem is now much better suited for numerical solution methods. It has been shown that, depending on the specific problem to be solved, the use of simplified kinetic models allows the computational time to be reduced by two to three orders of magnitude [85].

5.2. Reacting Channel Flows

In chemical microprocess technology one class of reactor designs deserves the term "generic", since many of the microreactors reported in the literature are based on this design concept. The design comprises at least one rectangular microchannel, often a multitude thereof, with a solid catalyst attached to the channel walls. The reacting fluid flows through the channel, while the reagents diffuse to the channel wall where they undergo chemical reactions. There are two versions of this design concept (Fig. 20). Either a smooth surface, often a metal layer, acts as catalyst, or the reaction occurs in a catalytically active porous medium. Clearly, the advantage of the porous catalyst layer is the higher specific surface area offering more reaction sites to the reagents. However, fully coupled simulations of reaction – convection – diffusion processes in catalyst-coated channels are quite rare, so most of the studies reported in this section are based on the concept of wall-catalyzed reactions.

For the reasons described above, reaction – convection – diffusion problems tend to be difficult to solve numerically. Hence, the simulation of reacting flows in three dimensions or parameter studies of micro reaction devices may be very time consuming. To enable rapid prototyping of microreactors, efficient modeling strategies with a minimum expenditure of computational resources are needed. The modeling approach developed in [86] allows a limited class of reacting microchannel flows to be assessed very quickly. It assumes a microchannel of length L and depth h with a first-order reaction occurring at one of the channel walls (Fig. 21). In cases where the flow profile is independent of the axial position in the channel and the problem can be approximated by a two-dimensional model, the mass transport equation for a chemical species a can be written in dimensionless form as

$$\frac{\mathrm{Pe}\,h}{L}u\left(\eta\right)\frac{\partial c_\mathrm{a}}{\partial\zeta}=\frac{\partial^2 c_\mathrm{a}}{\partial\eta^2} \tag{39}$$

where the axial and transverse coordinates ζ and η were nondimensionalized by the channel length L and the channel depth h. The reactant concentration is denoted by c_a, the velocity by u, and the Péclet number is expressed by the average velocity $\bar{u}$ and the diffusion constant D as $\mathrm{Pe}=\bar{u}\,h/D$. Equation 39 is solved subject to the boundary conditions of an impermeable upper channel wall and a first-order reaction with rate constant k occurring at the lower channel wall.

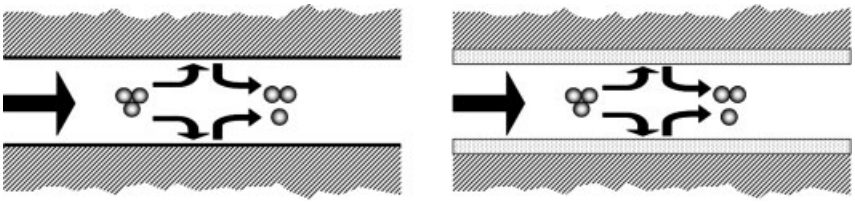

Figure 20. Reaction channels with a smooth surface (left) and a porous medium (right) as catalyst

Such a first-order reaction term surely does not adequately capture the mechanism of heterogeneous catalysis, but it may be a reasonable approximation of the kinetics in a limited parameter or operation range. An important dimensionless group characterising the reactive flow is the Damköhler number, defined as

$$\mathrm{Da}=\frac{kh}{D} \tag{40}$$

which characterizes the ratio of the diffusive and reactive timescales. The mass-transport equation has a separable solution of the form

$$c_\mathrm{a}\left(\zeta,\eta\right)=\bar{c}_\mathrm{a}\left(\zeta\right)f_\mathrm{a}\left(\eta\right), \tag{41}$$

where

$$\bar{c}_\mathrm{a}\left(\zeta\right)=\int_0^1 c_\mathrm{a}\left(\zeta,\eta\right)\mathrm{d}\eta. \tag{42}$$

By inserting this ansatz into Equation 39, the solution can be determined in form of an eigenfunction expansion [87]. The parameter controlling the number of terms of this expansion having to be taken into account is $Pe{\cdot}h/L$, which is usually of the order $0.01 - 1$ in microreactors. For this reason, often only the first term contributes. With the entrance condition $\bar{c}_\mathrm{a}\left(\zeta\right)=1$ the axial dependence can then be written as

$$\bar{c}_\mathrm{a}\left(\zeta\right)=\exp\left(-\lambda_\mathrm{a}\zeta\right), \tag{43}$$

where the eigenvalue λ_a is given as the solution of a nonlinear algebraic equation. Comparing the analytical results to full numerical simulations showed good agreement [86]. In addition to isothermal flows, analytical solutions were also determined for nonisothermal reacting flows and the model extended to second-order kinetics. Thus, a class of models was developed which may provide a simple characterization of reacting flows in microchannels without the need to do a full numerical simulation.

A similar analytical model for reacting flows in microchannels was used to assess the quality of simple plug-flow models which may be used to estimate reaction rate constants [88]. Microreactors lend themselves to measuring intrinsic rate constants of chemical reactions as, due to the short diffusion paths, heat and mass transfer limitations can be eliminated. The simplest way to deduce the rate constant k of a first-order heterogeneously catalyzed reaction at the walls of a tube is by assuming the reaction to occur in the volume of a plug-flow reactor. In this way the wall reaction is replaced by a pseudohomogeneous reaction and the velocity profile of the flow is ignored, which means that effectively a one-dimensional model is used. By measuring the inlet and outlet concentration of the reacting component, the rate constant is then obtained as

$$k=\frac{\bar{u}\,R}{2L}\ln\left(\frac{c_\mathrm{a}\left(\zeta=1\right)}{c_\mathrm{a}\left(\zeta=0\right)}\right). \tag{44}$$

where the notation is chosen similar to the previous paragraph, and L and R are the length and radius of the tube, respectively. Two effects are not taken into account by this expression. First, radial concentration gradients are ignored. Second, dispersion in the tube as discussed in Section 4.4 is neglected.

The one-dimensional model of reacting flows was extended to include Taylor – Aris dispersion [88] by considering an equation of the form

$$\frac{\mathrm{d}^2c_\mathrm{a}}{\mathrm{d}\zeta^2}-\frac{\mathrm{Pe}^*}{2}\frac{\mathrm{d}c_\mathrm{a}}{\mathrm{d}\zeta}-\beta c_\mathrm{a}=0, \tag{45}$$

where Pe^* is a modified Péclet number containing the Taylor – Aris dispersion constant instead of the diffusivity, and β a dimensionless parameter representing the pseudohomogeneous reaction. To study the influence of dispersion on chemical conversion, the solution of Equation 45 was compared to the solution of the corresponding two-dimensional problem, obtained in a similar way as sketched in the previous paragraph. It turned out that for a

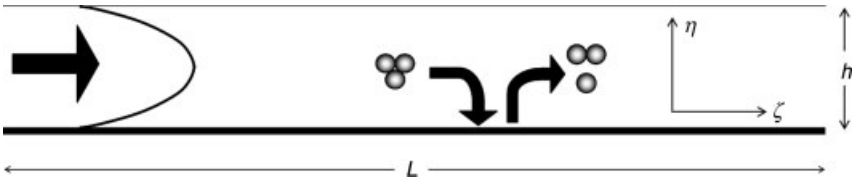

Figure 21. Two-dimensional model geometry of a microchannel with a reaction occurring at the lower channel wall

Damköhler number of 1, no satisfactory agreement between the one- and the two-dimensional models was achieved. The inclusion of Taylor – Aris dispersion improved the concentration profiles to a certain degree with respect to a plug-flow model, but the main reason for the deviations are the radial concentration gradients, which are not accounted for in the one-dimensional models. Hence, when attempting to extract intrinsic reaction-rate constants from comparisons of experimental results with results of one-dimensional reactor models, care should be taken to work in a regime of Damköhler numbers significantly smaller than 1.

5.3. Heat-Exchanger Reactors

The design of multichannel microreactors for gas-phase reactions is typically based on a stack of microstructured platelets. For strongly endothermic or exothermic reactions, it lends itself to alternate between layers of reaction channels and heating or cooling gas channels which supply energy to or withdraw it from the reaction. Such a set up is similar to the heat exchanger design depicted in Figure 9. Within this class of microreactor designs a choice can be made between different flow schemes of the gas streams in adjacent layers (co-, counter- or cross-current). The countercurrent coupling of an endothermic reaction to a heating gas stream in a multilayer architecture was studied [41] for a 2D model (Figure 22).

The dynamics of a heterogeneously catalyzed gas-phase reaction occurring in a nanoporous medium in combination with heat and mass transfer was simulated using a finite-volume approach. In contrast to other studies of similar type, heat and mass transfer in the nanoporous medium were explicitly accounted for by solving volume-averaged transport equations in the porous medium. Such an approach made it possible to compare the transport resistances in the gas phase and in the porous medium and to study the trade-off between maximization of catalyst mass and minimization of mass-transfer resistance due to pore diffusion. A typical concentration profile of a reacting chemical species which is converted by the catalyst is displayed in Figure 23. Due to the small pore size with an average diameter of 40 nm, the effective diffusivity in the porous medium is quite small and large concentration gradients build up, whereas in the microchannel the gradients are negligible. Typical catalyst effectiveness factors for a 100 µm catalyst layer were found to be of the order of 0.4. One of the outstanding potentials of microreactors is efficient utilization of the catalyst material. In conventional fixed-bed technology, catalyst pellets for liquid reactions are usually of a size of 2 – 5 mm [79]. Due to diffusive limitations in such comparatively large pellets, reactions often occur in a region close to the surface.

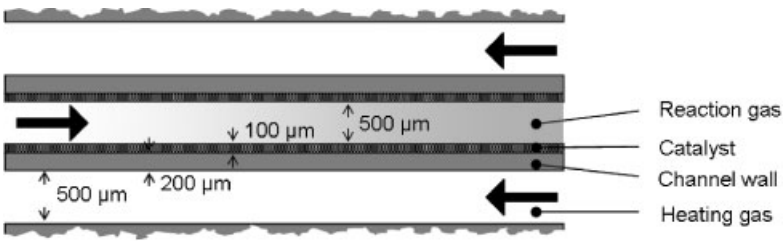

Figure 22. 2D model of a countercurrent heat-exchanger reactor with a nanoporous catalyst layer deposited on the channel wall

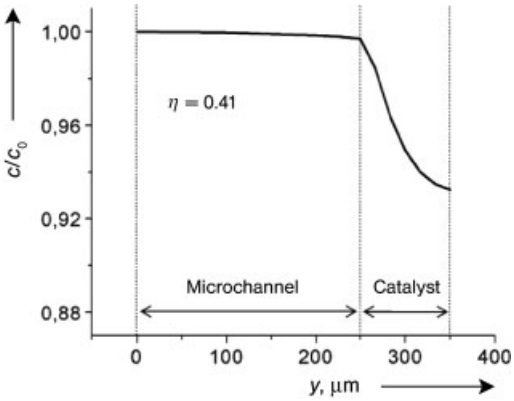

Figure 23. Normalized concentration profile of a reacting species across a microchannel of 500 µm width with a 100 µm catalyst layer deposited on the wall

A main objective of [41] was to study the influence of heat transfer on the achievable molar flux per unit reactor volume of the product species. Unstructured channels were compared to channels containing microfins. Heat-transfer enhancement due to microfins resulted in a different axial temperature profile with a higher outlet temperature in the reaction gas channel. Due to this effect and by virtue of the temperature dependence of the reaction rate, an improvement of heat transfer resulted in a significantly higher specific product molar flux. For the system under study, the heat-transfer enhancement achievable with microfins was found to increase the specific molar flux by about a factor of two. Such model studies show that a complex interplay between flow, heat transfer, and mass transfer may occur in microreactors and underline the need for fully coupled simulations incorporating conjugate heat transfer and transport in porous media.

The optimization of heat transfer in a heat-exchanger reactor was also the objective of [89]. Specifically, the exothermic water gas shift (WGS) reaction

$$CO + H_2O \leftrightarrow H_2 + CO_2 \tag{46}$$

which is utilized in fuel reformers to reduce the level of carbon monoxide, was considered. When the temperature of an exothermic, reversible reaction such as the WGS reaction increases, the kinetics are accelerated but the equilibrium is shifted more towards the feed components. Due to this fact, neither very low nor very high temperatures are optimal when the goal is to maximize the space – time yield for a given conversion. Rather, there is a specific temperature trajectory, i.e., a specific functional dependence of the reaction temperature on time, which allows maximization of the space – time yield. Due to their short thermal diffusion paths, microreactors allow the temperature profile in a reaction channel to be controlled much better than conventional equipment.

A counter-current heat-exchanger reactor for the WGS reaction with integrated cooling gas channels for removal of the reaction heat was analyzed [89]. The computational domain of the 2D model on the basis of the finite-element method is sketched in Figure 24. The reactor design does not allow for detailed adjustment of the temperature profile in the reaction gas channel, but by varying the cooling gas inlet temperature and the ratio of cooling gas and reaction gas flow rate, different temperature profiles can be imposed.

The reaction dynamics studied in the simulations is such that by release of reaction heat close to the inlet, the reaction gas temperature rises considerably. Owing to the fast kinetics at high temperatures, much of the CO conversion is already achieved in the inlet region. The remaining conversion is then mainly due to a reduction of the temperature and the corresponding shift in chemical equilibrium. This part of the process requires a large channel length. Thus, even minor differences in conversion achievable by improved temperature control may result in a considerable reduction of reactor size and the required amount of catalyst. When considering more advanced reactor designs allowing for fine-tuning of the temperature trajectory, CFD simulations are indispensable for performance optimization.

6. Conclusion and Outlook

The discussion of modeling and simulation techniques for microreactors shows that the toolbox available at present is quite diverse and goes well beyond the standard capabilities of CFD methods available in commercial solvers. In microreactors special methods needed for the modeling of noncontinuum physics play only a minor role, and most of the effects are described by the

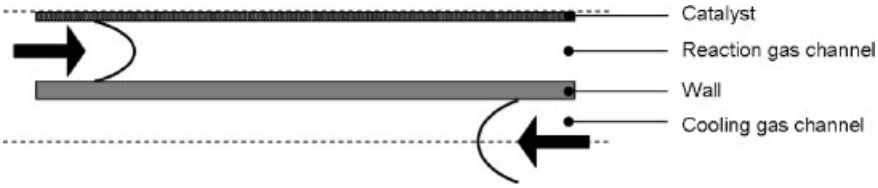

Figure 24. Model of a counter-current heat-exchanger reactor for exothermic reactions. The dashed lines indicate symmetry planes

standard continuum equations. However, even if the laminar nature of the flow somehow reduces the difficulty of simulation problems compared to macroscopic flows, there are a number of problems which are extremely difficult and require very fine computational grids. Among these problems is the numerical study of mixing in liquids, which often severely suffers from discretization artefacts.

A subject of growing importance which has not been discussed in this overview due to space limitations is modeling and simulation of multiphase microflows. Macroscopic multiphase flows are often modeled by using an Euler – Euler formulation of interpenetrating continua, that is, a volume-averaged description where phase boundaries are no longer explicitly resolved. When the spatial extension of bubbles or droplets becomes comparable to the channel dimensions, the gas – liquid or liquid – liquid interfaces must be modeled explicitly. There are methods available for such kinds of problems (e.g. the volume-of-fluid method) which are, however, computationally much more expensive than the Euler – Euler formulations employed for macroscopic flows. It is not only for the computational requirements that the predictive power of multiphase microflow computations is much lower than that of their counterparts for single-phase flow. As an example, moving contact lines are relevant for gas – liquid flows, and in some cases flow is accompanied by phase-change phenomena. Both the motion of a gas – liquid – solid contact line and the evaporation of liquid heavily depend on complex transport phenomena in the close vicinity of the contact-line region which are only partially understood until now. Consequently, the simulation of such phenomena as flow boiling in microchannels is in its infancy and its predictive power is very limited. Such examples underline the necessity of further research on multiphase microflows. Currently, corresponding numerical schemes are available, but the models do not adequately account for the boundary conditions at the contact line.

A further topic which has not been discussed in this overview and which is only beginning to evolve refers to system-level simulations of microprocess devices. In conventional process technology there is often an interplay between detailed CFD and system-level simulations of physicochemical processes. Owing to their large computational requirements, CFD methods can usually not be employed for the simulation of chemical reactors or plants consisting of a number of different subunits. Rather, CFD results enter into system models which treat the subunits in a simplified manner by just referring to characteristic curves specifying input – output behavior. In such a way, the flow in a pipe, for example, is characterized by a flow rate as a function of applied pressure difference. In conventional process technology a library of component models already exists which may be further extended and refined by incorporating the results of CFD simulations. However, in microprocess technology such system-level approaches do not yet seem to have been developed to an appreciable degree. A natural step for the further development of modeling and simulation techniques for microreactors would be to formulate system-level models which allow coupling to CFD simulations. In such a way also the commercial application of microreactors could gain additional impetus, as system-level models would allow fast evaluation of processes unparalleled by the comparatively time-consuming approaches presently available.

7. References

1. S. B. Pope, *Turbulent Flows*, Cambridge University Press, Cambridge 2000.
2. M. Gad-el-Hak, "The Fluid Mechanics of Microdevices", *J. Fluids Eng.* **121** (1999) 5 – 33.
3. C. Cercignani, *The Boltzmann Equation and its Applications*, Springer-Verlag, New York 1988.
4. G. Bird, *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Oxford University Press, New York 1994.
5. J. Pfahler, J. Harley, H. Bau, H. J. Zemel: "Liquid Transport in Micron and Submicron Channels", *Sensors Actuators A* **21 – 23** (1990) 431 – 434.
6. X. F. Peng, G. P. Peterson, B. X. Wang: "Frictional Flow Characteristics of Water Flowing through Rectangular Microchannels", *Experimental Heat Transfer* **7** (1994) 249 – 264.
7. X. F. Peng, G. P. Peterson: "The Effect of Thermofluid and Geometrical Parameters on Convection of Liquids through Rectangular Microchannels", *Int. J. Heat Mass Transfer* **38** (1995) 755 – 758.
8. S. M. Flockhart, R. S. Dhariwal: "Experimental and Numerical Investigation into the Flow Characteristics of Channels Etched in (100) Silicon", *J. Fluids Eng.* **120** (1998) 291 – 295.
9. T. M. Harms, M. J. Kazmierczak, F. M. Gerner: "Developing Convective Heat Transfer in Deep Rectangular Microchannels", *Int. J. Heat Fluid Flow* **20** (1999) 149 – 157.
10. G. M. Mala, D. Li: "Flow Characteristics of Water in Microtubes", *Int. J. Heat Fluid Flow* **20** (1999) 142 – 148.
11. W. Qu, G. M. Mala, D. Li: "Pressure-Driven Water Flows in Trapezoidal Silicon Microchannels", *Int. J. Heat Mass Transfer* **43** (2000) 353 – 364.
12. J. Baudry, E. Charlaix, A. Tonck, D. Mazuyer: "Experimental Evidence for a Large Slip Effect at a Nonwetting Fluid-Solid Interface", *Langmuir* **17** (2001) 5232 – 5236.
13. V. S. J. Craig, C. Neto, D. R. M. Williams: "Shear-Dependent Boundary Slip in an Aqueous Newtonian Liquid", *Phys. Rev. Lett.* **87** (2001) 54504 – 54507.
14. E. Bonaccurso, M. Kappl, H. J. Butt: "Hydrodynamic Force Measurements: Boundary Slip of Water on Hydrophilic Surfaces and Electrokinetic Effects", *Phys. Rev. Lett.* **88** (2002) 76103 – 76106.
15. M. Rieth, W. Schommers (eds.): *Theoretical and Computational Nanotechnology*, American Scientific Publishers, Stevenson Ranch 2005.
16. P. A. Thompson, S. M. Troian: "A General Boundary Condition for Liquid Flow at Solid Surfaces", *Nature* **389** (1997) 360 – 362.
17. R. F. Probstein, *Physicochemical Hydrodynamics*, Wiley, New York 1994.
18. R. J. Hunter, *Zeta Potential in Colloid Science: Principles and Applications*, Academic Press, New York 1981.
19. C. Yang, D. Li, J. H. Masliyah: "Modeling Forced Liquid Convection in Rectangular Microchannels with Electrokinetic Effects", *Int. J. Heat Mass Transfer* **41** (1998) 4229 – 4249.
20. G. E. Karniadakis, A. Beskok: *Micro Flows — Fundamentals and Simulation*, Springer-Verlag, New York 2002.
21. V. Hessel, S. Hardt, H. Löwe: *Chemical Micro Process Engineering*, Wiley-VCH, Weinheim 2004.
22. J. H. Ferziger, M. Peric: *Computational Methods for Fluid Dynamics*, 3rd ed., Springer-Verlag, Berlin 2002.
23. J. P. Hartnett, M. Kostic: "Heat Transfer to Newtonian and non-Newtonian Fluids in Rectangular Ducts", *Adv. Heat Transfer* **19** (1989) 247 – 356.
24. H. F. P. Purday: *An Introduction to the Mechanics of Viscous Flow*, Dover, New York 1949
25. R. K. Shah, A. L. London: "Laminar Flow Forced Convection in Ducts", *Adv. Heat Transfer* Suppl. 1, 1978.
26. R. K. Shah: "Laminar Flow Friction and Forced Convection Heat Transfer in Ducts of Arbitrary Geometry", *Int. J. Heat Mass Transfer* **18** (1975) 849 – 842.
27. D. H. Richardson, D. P. Sekulic, A. Campo: "Low Reynolds Number Flow Inside Straight Microchannels with Irregular Cross Sections", *Heat Mass Transfer* **36** (2000) 187 – 193.
28. Y. Asako, M. Faghri: "Finite-Volume Solutions for Laminar Flow and Heat Transfer in a Corrugated Duct", *J. Heat Transfer* **109** (1987) 627 – 634.
29. R. C. Xin, W. Q. Tao: "Numerical Prediction of Laminar Flow and Heat Transfer in Wavy Channels of Uniform Cross-Sectional Area", *Numerical Heat Transfer* **14** (1988) 465 – 481.

30. V. K. Garg, P. K. Maji: "Flow and Heat Transfer in a Sinusoidally Curved Channel", *Int. J. Eng. Fluid Mech.* **1** (1988) 293 – 319.
31. A. M. Guzmán, C. H. Amon: "Dynamical Flow Characterization of Transitional and Chaotic Regimes in Converging-Diverging Channels", *Int. J. Eng. Fluid Mech.* **321** (1996) 25 – 57.
32. W. R. Dean: "Note on the Motion of a Fluid in a Curved Pipe", *Philos. Mag.* **4** (1927) 208 – 223.
33. W. R. Dean: "The Stream-line Motion of Fluid in a Curved Pipe", *Philos. Mag.* **5** (1928) 673 – 695.
34. L. Wang, T. Yang: "Multiplicity and Stability of Convection in Curved Ducts: Review and Progress", *Adv. Heat Transfer* **38** (2004) 203 – 254.
35. Y. Chen, P. Chen: "Heat Transfer and Pressure Drop in Fractal Tree-Like Microchannel Nets", *Int. J. Heat Mass Transfer* **45** (2002) 2643 – 2648.
36. V. Cominos et al.: "A Methanol Steam Micro-Reformer for Low Power Fuel Cell Applications", *Chem. Eng. Comm.* **192** (2005) 685 – 698.
37. J. M. Commenge, L. Falk, J. P. Corriou, M. Matlosz: "Optimal Design for Flow Uniformity in Microchannel Reactors", *AIChE J.* **48** (2000) 345 – 358.
38. M. Kaviany: *Principles of Heat Transfer in Porous Media*, Springer-Verlag, New York 1995.
39. S. J. Kim, D. Kim, D. Y. Lee: "On the Local Thermal Equilibrium in Microchannel Heat Sinks", *Int. J. Heat Mass Transfer* **43** (2000) 1735 – 1748.
40. R. C. Xin, W. Q. Tao: "Numerical Predictions of Laminar Flows and Heat Transfer in Wavy Channels of Uniform Cross Section", *Numerical Heat Transfer* **14** (1988) 465 – 481.
41. S. Hardt, W. Ehrfeld, V. Hessel, K. M. van den Bussche: "Strategies for Size Reduction of Microreactors by Heat Transfer Enhancement Effects", *Chem. Eng. Comm.* **190** (2003) 540 – 559.
42. T. A. Rush, T. A. Newell, A. M. Jacobi: "An Experimental Study of Flow and Heat Transfer in Sinusoidal Wavy Passages", *Int. J. Heat Mass Transfer* **42** (1999) 1541 – 1553.
43. G. Wang, S. P. Wanka: "Convective Heat Transfer in Periodic Wavy Passages", *Int. J. Heat Mass Transfer* **38** (1995) 3219 – 3230.
44. D. R. Sawyers, M. Sen, H. C. Chang: "Heat Transfer Enhancement in Three-Dimensional Corrugated Channel Flow", *Int. J. Heat Mass Transfer* **41** (1998) 3559 – 3573.
45. C. C. Wang, C. K. Chen: "Forced Convection in a Wavy Wall Channel", *Int. J. Heat Mass Transfer* **45** (2002) 2587 – 2595.
46. S. Hardt, T. Baier: "A Computational Model for Heat Transfer in Multichannel Microreactors", Proc. of ICMM05 - 3rd Int. Conf. on Microchannels and Minichannels, June 13–15, 2005, Toronto, Canada, paper ICMM 2005-75079.
47. T. Stief, O.-U. Langer, K. Schubert: "Numerical Investigations of Opimal Heat Conductivity in Micro Heat Exchangers", *Chem. Eng. Technol.* **22** (1999) 297 – 302.
48. R. B. Peterson: "Numerical Modeling of Conduction Effects in Microscale Counterflow Heat Exchangers", *Microscale Thermophys. Eng.* **3** (1999) 17 – 30.
49. G. Maranzana, I. Perry, D. Maillet: "Mini- and Microchannels: Influence of Axial Conduction in the Walls", *Int. J. Heat Mass Transfer* **47** (2004) 3993 – 4004.
50. V. Hessel, H. Löwe, F. Schönfeld: "Micromixers — A Review on Passive and Active Mixing Principles", *Chem. Eng. Sci.* **60** (2005) 2479 – 2501.
51. N. T. Nguyen, Z. Wu, "Micromixers — A Review", *J. Micromech. Microeng.* **15** (2005) R1 – R16.
52. S. Hardt, K. S. Drese, V. Hessel, F. Schönfeld: "Passive Micromixers for Applications in the Microreactor and μTAS Fields", *Microfluidics Nanofluidics* **1** (2005) 108 – 118.
53. D. Gobby, P. Angeli, A. Gavriilidis: "Mixing Characteristics of T-Type Microfluidic Mixers", *J. Micromech. Microeng.* **11** (2001) 126 – 132.
54. D. Bothe, C. Stemich, H. J. Warnecke: "Theoretische und experimentelle Untersuchungen der Mischvorgänge in T-förmigen Mikroreaktoren · Teil 1: Numerische Simulation und Beurteilung des Strömungsmischens", *Chem. Ing. Tech.* **76** (2004) 1480 – 1484.
55. J. M. Ottino: *The Kinematics of Mixing: Stretching, Chaos and Transport*, Cambridge University Press, Cambridge 1989.
56. C. A. J. Fletcher: *Computational Techniques for Fluid Dynamics*, 2nd ed. Springer-Verlag, Berlin 1991.
57. B. Noll: *Numerische Strömungsmechanik*, Springer-Verlag, Berlin 1993.

58. A. D. Stroock et al.: "Chaotic Mixer for Microchannels", *Science* **295** (2002) 647 – 651.
59. H. Aref: "Stirring by Chaotic Advection", *Int. J. Eng. Fluid Mech.* **143** (1984) 1 – 21.
60. T. G. Kang, T. H. Kwon: "Colored Particle Tracking Method for Mixing Analysis of Chaotic Micromixers", *J. Micromech. Microeng.* **14** (2004) 891 – 899.
61. M. A. Stremler, F. R. Haselton, H. Aref: "Designing for Chaos: Applications for Chaotic Advection at the Microscale", *Phil. Trans. R. Soc. Lond. A* **362** (2004) 1019 – 1036.
62. R. H. Liu et al.: "Passive Mixing in a Three-Dimensional Serpentine Microchannel", *J. Microelectromech. Syst.* **9** (2000) 190 – 197.
63. X. Niu, Y. K. Lee: "Efficient Spatial-Temporal Chaotic Mixing in Microchannels", *J. Micromech. Microeng.* **13** (2003) 454 – 462.
64. P. Tabeling, M. Chabert, A. Dodge, C. Jullien, F. Okkels: "Chaotic Mixing in Cross-channel Micromixers", *Phil. Trans. R. Soc. Lond. A* **362** (2004) 987 – 1000.
65. F. Bottausci, I. Mezic, C. D. Meinhart, C. Cardonne: "Mixing in the Shear Superposition Micromixer: Three-Dimensional Analysis", *Phil. Trans. R. Soc. London A* **362** (2004) 1001 – 1018.
66. F. Jiang, K. S. Drese, S. Hardt, M. Küpper, F. Schönfeld: "Helical Flows and Chaotic Mixing in Curved Microchannels", *AIChE J.* **50** (2004) 2297 – 2305.
67. S. Hardt, H. Pennemann, F. Schönfeld: "Theoretical and Experimental Characterization of a Low Reynolds Number Split-And-Recombine Mixer", *Microfluidics Nanofluidics* **2** (2006) 237 – 248.
68. K. S. Drese: "Optimization of Interdigital Micromixers via Analytical Modeling — Exemplified with the SuperFocus Mixer", *Chem. Eng. J.* **101** (2004) 403 – 407.
69. S. Hardt, F. Schönfeld: "Laminar Mixing in Interdigital Micromixers with Different Mixing Chambers — Part 2: Numerical Simulations", *AIChE J.* **49** (2003) 578 – 584.
70. F. Schönfeld, V. Hessel, C. Hofmann: "An Optimised Split-and-Recombine Micro-Mixer with Uniform "Chaotic" Mixing", *Lab on a Chip* **4** (2004) 65 – 69.
71. H. Chen, J. C. Meiners: "Topologic Mixing on a Microfluidic Chip", *Appl. Phys. Lett.* **84** (2004) 2193 – 2195.
72. G. I. Taylor: "Dispersion of Soluble Matter in Solvent Flowing Slowly through a Tube", *Proc. Roy. Soc. A* **219** (1953) 186 – 203.
73. R. Aris: "On the Dispersion of a Solute in a Fluid Flowing through a Tube", *Proc. Roy. Soc. A* **235** (1956) 67 – 77.
74. W. N. Gill, R. Sankarasubramanian: "Exact Analysis of Unsteady Convective Diffusion", *Proc. Roy. Soc. Lond. Ser. A* **316** (1970) 341 – 350.
75. M. R. Doshi, P. M. Daiya, W. N. Gill: "Three Dimensional Laminar Dispersion in Open and Closed Rectangular Conduits", *Chem. Engn. Sci.* **33** (1978) 795 – 804.
76. R. Aris: "On the Dispersion of a Solute by Diffusion, Convection and Exchange between Phases", *Proc. R. Soc. London Ser. A* **252** (1959) 538 – 550.
77. D. Dutta, D. T. Leighton: "Dispersion Reduction in Pressure-Driven Flow through Microetched Channels", *Anal. Chem.* **75** (2001) 57 – 70.
78. P. Daskopoulos, A. M. Lenhoff: "Dispersion Coefficient for Laminar Flow in Curved Tubes", *AIChE J.* **34** (1988) 2052 – 2058.
79. R. H. Perry, D. W. Green: *Perry's Chemical Engineers' Handbook*, 7th ed. McGraw-Hill, New York 1997.
80. J. M. Thomas, W. J. Thomas: *Principles and Practice of Heterogeneous Catalysis*, VCH, Weinheim 1997.
81. W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery: *Numerical Recipes in Fortran 77*, Cambridge University Press, Cambridge 1992.
82. U. Maas: "Efficient Calculation of Intrinsic Low-Dimensional Manifolds for the Simplification of Chemical Kinetics", *Comput. Visualization Sci.* **1** (1998) 69 – 82.
83. U. Maas, S. B. Pope: "Simplifying Chemical Kinetics: Intrinsic Low-Dimensional Manifolds in Composition Space", *Combustion Flame* **88** (1992) 239 – 264.
84. M. Kiehl: "Partitioning Methods for the Simulation of Fast Reactions", *Zentralblatt Angew. Math. Mech.* **78** (1998) 967 – 970.
85. S. B. Pope, "Computationally Efficient Implementation of Combustion Chemistry using In Situ Adaptive Tabulation", *Combustion Theory Modeling* **1** (1997) 41 – 63.
86. D. Gobby, I. Eames, A. Gavriilidis: "A Vertically-Averaged Formulation of Catalytic Reactions in Microchannel Flows", Proc. of IMRET 5: 5th Int. Conf. on Microreaction

Technology, Springer-Verlag, Berlin 2001, pp. 141 – 149.
87. R. E. Walker: "Chemical Reaction and Diffusion in a Catalytic Tubular Reactor", *Phys. Fluids* **4** (1961) 1211 – 1216.
88. J. M. Commenge, L. Falk, J. P. Corriou, M. Matlosz: "Microchannel Reactors for Kinetic Measurement: Influence of Diffusion and Dispersion on Experimental Accuracy", Proc. of IMRET 5: 5th Int. Conf. on Microreaction Technology, Springer-Verlag, Berlin 2001, pp. 131 – 140.
89. W. E. TeGrotenhuis, D. L. King, K. P. J. Brooks, G. B. R. S. Wegeng: "Optimizing Microchannel Reactors by Trading-Off Equilibrium and Reaction Kinetics through Temperature Management", Proc. of IMRET6: 6th Int. Conf. on Microreaction Technology, AIChE Pub. No. 164 (2002) 18 – 28.

Benzofurans

GERD COLLIN, Rütgerswerke AG, Duisburg, Germany
HARTMUT HÖKE, Rütgerswerke AG, Duisburg, Germany

1. Benzofuran 1
2. Dibenzofuran 1
3. References 3

1. Benzofuran

Benzofuran [*271-89-6*], coumarone, C_8H_6O, was obtained by R. FITTIG and G. EBERT in 1882, by calcining coumarilic acid (from coumarin) with lime. In 1890, G. KRAEMER and A. SPILKER discovered benzofuran in coal tar.

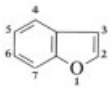

Physical Properties. Benzofuran, M_r 118.14, *bp* (101.3 kPa) 171.4 °C, *mp* −28.9 °C, ϱ (20 °C) 1.0948 g/cm^3, is a colorless liquid with an aromatic odor; it is volatile in steam, soluble in ethanol, and insoluble in water.

Chemical Properties. Benzofuran polymerizes slowly at ambient temperature, more rapidly with heat and in the presence of an acidic catalyst; it is resistant to aqueous alkali.

Production. Coal tar contains ca. 0.01 % benzofuran, which can be concentrated through distillation to ca. 15 % in a light oil fraction with a boiling range of 165–175 °C. It is recovered either by precipitation as the dibromide or the picrate, or by sulfonation in acetic anhydride, followed by decomposition of the benzofuran derivatives. If the benzofuran fraction is largely free from indene, benzofuran can be concentrated further through either selective extraction or azeotropic distillation with alkylene glycols.

Benzofuran can be synthesized by dehydrogenation and cyclization of 2-ethylphenol.

Uses. Benzofuran, together with indene, is the raw material used for the production of indene–coumarone resins (→) from light oil fractions of coal tar. Normal commercial resins contain up to 10 % benzofuran polymers. No large-scale commercial use has been found for pure benzofuran.

Toxicology and Occupational Health. Benzofuran showed acute toxic effects in rabbits when administered orally at a dosage of 1 g [8]. The LD_{50} (mice, intraperitoneal) is 500 mg/kg [9]. (A carcinogenicity study on benzofuran in rats and mice was performed in the United States under the National Toxicology Program. The results have not yet been published [10].)

2. Dibenzofuran

Dibenzofuran [*132-64-9*], diphenylene oxide, $C_{12}H_8O$, was synthesized for the first time by C. LESIMPLE in 1866 from triphenyl phosphate and lime; its constitution was recognized in 1871 by W. HOFFMEISTER, and it was found in coal tar by G. KRAEMER and R. WEISSGERBER in 1901.

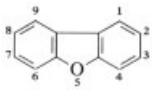

Physical Properties. Dibenzofuran, M_r 168.18, *bp* (101.3 kPa) 287 °C, ϱ(99.3 °C) 1.0886 g/cm^3, *mp* 82.8 °C (a higher melting point indicates impurities of fluorene or acenaphthene), consists of colorless crystals with

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.l03_l01

a weak, characteristic odor. It is resistant for a short time in an inert atmosphere to >600 °C. Dibenzofuran is volatile in steam; sublimable; readily soluble in diethyl ether, benzene, and acetic acid; fairly soluble in ethanol; insoluble in water; its enthalpy of vaporization (130–140 °C) is 66.4 kJ/mol, enthalpy of combustion 5882 kJ/mol, dielectric constant (100 °C) 3.0, and dipole moment (liquid) 0.9 D.

Chemical Properties. The skeleton of the dibenzofuran molecule combines that of both biphenyl and diphenyl ether. In comparison with biphenyl, the reactivity of dibenzofuran to electrophilic substitution is lower. The 2- and 2,8-derivatives can be prepared by halogenation, sulfonation, Friedel–Crafts reaction, Gattermann's aldehyde synthesis, and chloromethylation; the 3- and 3,8-derivatives are obtained by nitration; the 4- and 4,6-positions exclusively by metalation and mercuration. In the 1-position, substituted derivatives can be formed only by indirect methods or by radical substitution. If dibenzofuran is already substituted, this substituent determines whether a new substituent is homonuclear, i.e., in the same ring (e.g., hydroxyl, amino, ether groups), or heteronuclear (e.g., halogen, nitro, acyl groups). Examples of opening the furan nucleus include fusion with potassium hydroxide at 280–300 °C to give 2,2′-dihydroxybiphenyl, reaction with potassium metal to form 2-hydroxybiphenyl, or conversion to phenol with aluminum chloride at 140 °C.

Production. Dibenzofuran is present in coal tar at a level of ca. 1 %. It is recovered from a wash oil fraction that boils between 275 and 290 °C (containing ca. 30 % dibenzofuran). A redistillation step is largely successful in separating dibenzofuran from acenaphthene, which boils 7 °C lower. The technically pure compound is obtained through crystallization of the redistilled fraction. In admixture with ca. 60 % acenaphthene, dibenzofuran forms a eutectic and, with fluorene, which boils about 13 °C higher, a continuous series of mixed crystals [11].

Dibenzofuran can be synthesized by dehydrogenation of phenol at 450 °C [12]; by oxidative dehydrogenation of phenol and cyclization of 2-cyclohexenylcyclohexanone [13]; by cyclization of diphenyl ether in the presence of palladium(II) acetate [14]; or by the pyrolysis of phthalic anhydride with furan [15]. As sufficient quantities of dibenzofuran can be produced from coal tar, these syntheses are not industrially applied. Some dibenzofuran derivatives can be synthesized better by ring closure than by substitution of dibenzofuran.

Uses. Due to its high heat resistance, dibenzofuran, together with biphenyl (→), is a component of heat-transfer oils. When combined with methylnaphthalenes, it is suitable as a carrier for dyeing and printing textiles [16]. Some of its derivatives serve as intermediates for naphtol AS (→) and chromium complex dyes, e.g., 2,2′-dihydroxybiphenyl (→) and 2-hydroxydibenzofuran-3-carboxylic acid (from 2-hydroxydibenzofuran through Kolbe–Schmitt carboxylation).

Derivatives of dibenzofuran phosphonic acid can be used as antioxidants in plastics [17]. A range of polymers can be produced from dibenzofuran, e.g., heat-resistant polyarylacetylene [18] and quinoxaline polymers [19] or photoconductive polymers for electrophotography [20]. Dibenzofuroyl silicon polymers and monomers can be employed as perfume oil bases, in cosmetics, and as fluids for the production of elastomers and resins [21]. Dibenzofuran derivatives have been proposed as pharmaceuticals, e.g., dibenzofuranyl aminoalcohols as sedatives and tranquilizers [22] and dibenzofurylpropionic acid and derivatives as antiphlogistics [23].

Toxicology and Occupational Health. Dibenzofuran caused no adverse effects when fed to rats at high dosage for 200 d (0.025–0.4 % of feed) [24]. In skin- and eye-irritant tests (24-h exposure, FDA method), dibenzofuran proved to be nonirritating [25]. On exposure of fish (mullet sp.) to 750 μg/L for 6 and 24 d, dibenzofuran caused weak biochemical stress responses in brain, interrenal gland, and liver. This dosage was one-third the one-week LC_{50} level [26, 27]. Dibenzofuran is degraded by microbes [28, 29]. The microbial fate of benzofuran is unknown [30], although it is metabolized in mammals (rats) by oxidation, followed by conjugation with sulfate and excretion in urine [8]. No TLV has been stipulated for either substance.

3. References

General References

1. *Beilstein*, **17**, 272–278;
   17(1), 145–147;
   17(2), 305–309;
   17(3/4), 4403–4412;
   17(5), 292–299.
2. *Beilstein*, **17**, 54;
   17(1), 24;
   17(2), 57;
   17(3/4), 478–480.
3. *Beilstein*, **17**, 70;
   17(1), 30;
   17(2), 67;
   17(3/4), 585–586.
4. H.-G. Franck, G. Collin: *Steinkohlenteer*, Springer Verlag, Berlin–Heidelberg–New York 1968, pp. 11, 41, 93.
5. A. Mustafa: *Benzofurans. Chemistry of Heterocyclic Compounds (Weissberger)*, vol. **29**, Wiley Interscience, New York 1974.
6. P. Cagniant, D. Cagniant in A. R. Katritzky, A. J. Boulton (eds.) *Advances in Heterocyclic Chemistry*, vol. **18**, Academic Press, New York–San Francisco–London 1975, pp. 337–482.
7. M. V. Sargent, P. O. Stransky in A. R. Katritzky (ed.): *Advances in Heterocyclic Chemistry*, vol. **35**, Academic Press, Orlando 1984, pp. 2–81.

Specific Reference

8. F. Böhm, *Hoppe-Seyler's Z. Physiol. Chem.* **269** (1941) 17–23.
9. M. Fatome, L. Andrieu, J.-D. Laval, J.-M. Clavel et al. *Eur. J. Med. Chem.-Chim. Ther.* **12** (1977) 383–384.
10. U.S. Department of Health & Human Services: personal communication.
11. G.-P. Blümer, G. Collin, *Erdöl Kohle Erdgas Petrochem*. **36** (1983) 22–27.
12. Monsanto, US 4 008 254 1975 (D. E. Gross, N. A. Fishel); US 4 013 694 (N. A. Fishel).
13. Anic SpA., DE 2 552 652, 1975 (P. A. Moggi, G. Iori).
14. Schering, DE 2 418 503, 1974 (B. Akermark, L. Eberson, E. Jonsson, E. Pettersson); Halcon SD Group Inc., US 4 362 883, 1981 (R. J. Harvey); Ashland Oil Inc., US 4 490 550, 1983 (A. B. Goel).
15. Standard Oil, US 3 514 458, 1967 (E. K. Fields).
16. Cassella Farbwerke Mainkur, Rütgerswerke, DE 2 236 551, 1972 (G. Weckler, R. Mildenberg).
17. Sandoz, DE 2 215 544, 1972 (K. Hofer, G. Tscheulin).
18. Hercules Inc., US 4 026 859, 1975 (L. C. Cessna).
19. US Dept. of the Air Force, US 3 792 017, 1972 (F. E. Arnold, F. L. Hedberg, R. F. Kovar).
20. Azoplate Corp., US 3 294 531, 1960 (H. Schlesinger).
21. General Electric, US 3 715 370, 1970; US 3 855 241, 1972 (E. V. Wilkus, A. Berger).
22. Aldrich Chemical Co., US 3 701 786, US 3 716 539, US 3 751 390, 1971 (H. B. Hopps, D. Jackman, J. H. Biel).
23. Merck Patent GmbH, DE 2 261 745, 1972 (J. Gante, W. Mehrhof, A. Wild).
24. J. O. Thomas, R. H. Wilson, C. W. Eddy, *Food Res*. **5** (1940) 23–30.
25. H. Willeitner, H. O. Dieter, *Holz Roh-Werkst*. **42** (1984) 223–232.
26. P. Thomas, H. W. Wofford, J. M. Neff, *Aquat. Toxicol*. **1** (1981) 329–342.
27. P. Thomas, H. W. Wofford, Toxicol. *Appl. Pharmacol*. **76** (1984) 172–182.
28. M. D. Lee, J. T. Wilson, C. H. Ward, *Dev. Ind. Microbiol*. **25** (1984) 557–565.
29. C. E. Cerniglia, J. C. Morgan, D. T. Gibson, *Biochem. J.* **180** (1979) 175–185.
30. P. W. Trudgill in D. T. Gibson (ed.): *Microbial Degradation of Organic Compounds*, Marcel Dekker, New York 1984, pp. 295–308.

Biorefineries – Industrial Processes and Products

Birgit Kamm, Research Institute Bioactive Polymer Systems (biopos e.V.), Teltow, Germany
Patrick R. Gruber, Outlast Technologies Incorporated, Boulder, Colorado, USA
Michael Kamm, Biorefinery.de GmbH, Potsdam, Germany

1. Introduction 1
2. Historical Outline 2
2.1. Historical Technological Outline and Industrial Resources 2
2.2. The Beginning – A Digest 2
2.2.1. Sugar Production 2
2.2.2. Starch Hydrolysis 2
2.2.3. Wood Saccharification 3
2.2.4. Furfural 3
2.2.5. Cellulose and Pulp 3
2.2.6. Levulinic Acid 3
2.2.7. Lipids 4
2.2.8. Vanillin from Lignin 4
2.2.9. Lactic Acid 4
2.3. The Origins of Integrated Biobased Production 4
3. Situation 7
3.1. Some Current Aspects of Biorefinery Research and Development 7
3.2. Raw Material Biomass 7
3.3. National Vision and Goals and Plan for Biomass Technology in the United States 9
3.4. Vision and Goals and Plan for Biomass Technology in the European Union and Germany . . . 9
4. Principles of Biorefineries 10
4.1. Fundamentals 10
4.2. Definition of the Term "Biorefinery" 12
4.3. The Role of Biotechnology 13
4.3.1. Guidelines of Fermentation Section within Glucose-product Family Tree . 13
4.4. Building Blocks, Chemicals and Potential Screening 15
5. Biorefinery Systems and Design . . 16
5.1. Introduction 16
5.2. Lignocellulosic Feedstock Biorefinery 16
5.3. Whole-crop Biorefinery 18
5.4. Green Biorefinery 20
5.5. Two-platform Concept and Syngas 21
6. Biorefinery Economy 23
7. Outlook and Perspectives 25
8. References 26

1. Introduction

The preservation and management of our diverse resources are fundamental political tasks to foster sustainable development in the 21st century. Sustainable economic growth requires safe and sustainable resources for industrial production, a long-term and confident investment and finance system, ecological safety, and sustainable life and work perspectives for the public. Fossil resources are not regarded as sustainable, however, and their availability is more than questionable in the long-term. Because of the increasing price of fossil resources, moreover, the feasibility of their utilization is declining.

It is, therefore, essential to establish solutions which reduce the rapid consumption of fossil resources, which are not renewable (petroleum, natural gas, coal, minerals). A forward looking approach is the stepwise conversion of large parts of the global economy into a sustainable biobased economy with bioenergy, biofuels, and biobased products as its main pillars (Fig. 1).

Whereas for energy production a variety of alternative raw materials (wind, sun, water, biomass, nuclear fission and fusion) can be established, industry based on conversion of sustainable material, for example the chemical industry, industrial biotechnology, and also the fuel generation, depends on biomass, in particular mainly on plant biomass.

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.l04_l01

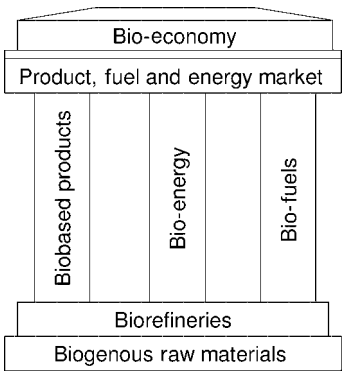

Figure 1. 3-Pillar model of a future biobased economy

Some change from the today's production of goods and services from fossil to biological raw materials will be essential. The rearrangement of whole economies to implement biological raw materials as a source with increased value requires completely new approaches in research and development. On the one hand, biological and chemical sciences will play a leading role in the generation of future industries in the 21st century. On the other hand, new synergies of biological, physical, chemical, and technical sciences must be elaborated and established. This will be combined with new traffic technology, media- and information technology, and economic and social sciences. Special requirements will be placed on both the substantial converting industry and research and development with regard to raw material and product line efficiency and sustainability.

The development of substance-converting basic product systems and polyproduct systems, for example biorefineries, will be the "key for the access to an integrated production of food, feed, chemicals, materials, goods, and fuels of the future" [4].

2. Historical Outline

2.1. Historical Technological Outline and Industrial Resources

Today's biorefinery technologies are based (1) on the utilization of the whole plant or complex biomass and (2) on integration of traditional and modern processes for utilization of biological raw materials. In the 19th and the beginning of the 20th century large-scale utilization of renewable resources was focused on pulp and paper production from wood, saccharification of wood, nitration of cellulose for guncotton and viscose silk, production of soluble cellulose for fibers, fat curing, and the production of furfural for Nylon. Furthermore, the technology of sugar refining, starch production, and oil milling, the separation of proteins as feed, and the extraction of chlorophyll for industrial use with alfalfa as raw material were of great historical importance. But also processes like wet grinding of crops and biotechnological processes like the production of ethanol, acetic acid, lactic acid, and citric acid used to be fundamental in the 19th and 20th century.

2.2. The Beginning – A Digest

2.2.1. Sugar Production

The history of industrial conversion of renewable resources is longer than 200 years. Utilization of sugar cane has been known in Asia since 6000 BC and imports of cane sugar from oversea plantations have been established since the 15th century. The German scientist A. S. Marggraf was a key initiator of the modern sugar industry. In 1748 he published his research on the isolation of crystalline sugar from different roots and beet [5, 6]. Marggraf's student, F. C. Achard, was the first to establish a sugar refinery based on sugar beet, in Cunern/Schlesien, Poland, in 1801.

2.2.2. Starch Hydrolysis

In 1811, the German pharmacist G. S. C. Kirchhoff found that when potato starch was cooked in dilute acid the starch was converted into "grape sugar" (i.e., *d*-glucose or dextrose) [7]. This was not only a very important scientific result but also the starting point of the starch industry. In 1806 the French emperor Napoleon Bonaparte introduced an economic continental blockade which considerably limited overseas trade in cane sugar. Thus, starch hydrolysis became of interest for the economy. The first starch sugar plant was established in Weimar, Germany, in 1812, because of a recommendation

of J. W. Döbereiner to grand duke Carl August von Sachsen-Weimar. Successful development of the sugar beet industry, however, initially obstructed further development of the starch industry [8]. In 1835, the Swedish Professor J. J. Berzelius developed enzymatic hydrolyses of starch into sugar and introduced the term "catalysis".

2.2.3. Wood Saccharification

In 1819 the French plant chemist H. Braconnot discovered that sugar (glucose) is formed by treatment of wood with concentrated sulfuric acid. 1855, G. F. Melsens reported that this conversion can be carried out with dilute acid also. Acid hydrolysis can be divided into two general approaches, based on (1) concentrated acid hydrolysis at low temperature and (2) dilute acid hydrolysis at high temperature. Historically, the first commercial processes, named wood saccharification, were developed in 1901 by A. Classen (Ger. Patent 130980), employing sulfuric acid, and in 1909 by M. Ewen and G. Tomlinson (US Patent 938208), working with dilute sulfuric acid. Several plants were in operation until the end of World War I. Yields of these processes were usually low, in the range 75–130 liter per ton wood dry matter only [9, 10]. Technologically viable processes were, however, developed in the years between World War I and World War II. The German chemist Friedrich Bergius was one of the developers. The sugar fractions generated by wood hydrolyses have a broad spectrum of application. An important fermentation product of wood sugar of increasing interest is ethanol. Ethanol can be used as fuel either blended with traditional hydrocarbon fuel or as pure ethanol. Ethanol is also an important platform chemical for further processing [11].

2.2.4. Furfural

Dübereiner was the first to report the formation and separation of furfural by distillation of bran with diluted sulfuric acid in 1831. In 1845 the English Chemist G. Fownes proposed the name "furfurol" (furfur–bran; oleum–oil). Later the suffix "ol" was changed to "al" because of the aldehyde function [12, 13].

Treatment of hemicellulose-rich raw materials with dry steam in the presence of hydrogen chloride gave especially good results [14]. Industrial technology for production of furfural from pentose is based on a development of an Anglo-American company named Quaker Oats. The process was been developed in the nineteen-twenties [E.P. 203691 (1923), F.P. 570531 (1923)]. Since 1922 Quaker Oats Cereal Mill in Cedar Rapids/Iowa, USA, has produced up to 2.5 tons of furfural per day from oat husks. Since 1934 the process had been established as an industrial furfural plant. Furfural was the cheapest aldehyde, at 16–17 cents per lb (lb=pound; 1 metric ton=1000 kg=2204.62442 lb; 1 kg=0.453592 lb) [15]. Until approximately 1960 DuPont used furfural as a precursor of nylon-6.6. Furfural has since been substituted by fossil based precursors.

2.2.5. Cellulose and Pulp

In 1839 the Frenchman A. Payen discovered that after treatment of wood with nitric acid and subsequent treatment with a sodium hydroxide solution a residue remained which he called "les cellules", cellulose [16]. In 1854 caustic soda and steam were used by the Frenchman M. A. C. Mellier to disintegrate cellulose pulp from straw. In 1863 the American B. C. Tilgham registered the first patent for production of cellulose by use of calcium bisulfite. Together with his brother, Tilgham started the first industrial experiments to produce pulp from wood by treatment with hydrogen sulfite. This was 1866 at the paper mill Harding and Sons, Manayunk, close to Philadelphia. In 1872 the Swedish Engineer C. D. Ekman was the first to produce sulfite cellulose by using magnesium sulfite as cooking agent [17]. By 1900 approximately 5200 pulp and paper mills existed worldwide, most in the USA, approximately 1300 in Germany, and 512 in France.

2.2.6. Levulinic Acid

In 1840 the Dutch Professor G. J. Mulder (who also introduced the name "protein") synthesized levulinic acid (4-oxopentanoic acid, c-ketovaleric acid) by heating fructose with hydrochloride for the first time. The former term

"levulose" for fructose gave the levulinic acid its name [18]. Although levulinic acid has been well known since the 1870s when many of its reactions (e.g., esters) were established, it has never reached commercial use in any significant volume. In the 1940s commercial levulinic acid production was begun in an autoclave in the United States by A. E. Staley, Dectur, Illinois [19]. At the same time utilization of hexoses from low-cost cellulose products was examined for the production of levulinic acid [20]. As early as 1956 levulinic acid was regarded as a platform chemical with high potential [21].

2.2.7. Lipids

From 1850 onward European import of tropical plant fats, for example palm-oil and coconut oil, started. Together with the soda process, invented by the French N. Leblanc in 1791, the industrialization of the soap production began and soap changed from luxury goods into consumer goods. The developing textile industry also demanded fat based products. In 1902 the German chemist W. Normann discovered that liquid plant oils are converting into tempered fat by augmentation of hydrogen. Using nickel as catalyst Norman produced tempered stearic acid by catalytic hydration of liquid fatty acids [22]. The so called "fat hardening" led to the use of European plant oils in the food industry (margarine) and other industries.

2.2.8. Vanillin from Lignin

In 1874 the German chemists W. Haarmann and F. Tiemann were the first to synthesize vanillin from the cambial juice of coniferous wood. In 1875 the company Haarmann and Reimer was founded. The first precursor for the production of vanillin was coniferin, the glucoside of coniferyl alcohol. This precursor of lignin made from cambial juice of coniferous trees was isolated, oxidized to glucovanillin, and then cleaved into glucose and vanillin [23]. This patented process [24] opened the way to industrial vanillin production. It was also the first industrial utilization of lignin. Besides the perfume industry, the invention was of great interest to the upcoming chocolate industry. Later, however, eugenol (1-allyl-4-hydroxy-3-methoxybenzene), isolated from clove oil, was used to produce vanillin. Today, vanillin production is based on lignosulfonic acid which is a side product of wood pulping. The lignosulfonic acid is oxidized with air under alkaline conditions [25, 26].

2.2.9. Lactic Acid

In 1895 industrial lactic acid fermentation has been developed by the pharmaceutical entrepreneur A. Boehringer. The Swedish pharmacist C. W. Scheele had already discovered lactic acid in 1780 and the conversion of carbohydrates into lactic acid had been known for ages in food preservation (e.g. Sauerkraut) or agriculture (silage fermentation). Because of the activity of Boehringer the German company Boehringer-Ingelheim can be regarded as the pioneer of industrial biotechnology. Both the process and the demand for lactic acid by dyeing factories, and the leather, textile, and food industries made the company the leading supplier. In 1932 W. H. Carothers, who was also the inventor of polyamide-6.6, developed, together with van Natta, a polyester made from lactic acid, poly(lactic acid) [27]. In the late 1990s this poly(lactic acid) was commercialized by the company NatureWorks (Cargill, the former Cargill Dow) [28].

2.3. The Origins of Integrated Biobased Production

In the year 1940 the German chemist P. von Walden (noted for his "Inversion of configuration at substitution reactions", the so-called "Walden-Reversion") calculated that in 1940 Germany produced $13 \cdot 10^6$ t of cellulose leaving 5 to 6 million tons of lignin suitable only as wastage. He then formulated the question: How long can national economy tolerate this [29]? Approaches to integrated production during industrial processing of renewable primary products have a long tradition, starting from the time when industrial cellulose production expanded continuously, as also did the related waste-products. Typical examples of this will be mentioned.

As early as 1878 A. Mitscherlich, a German chemist, started to improve the sulfite pulp process by fermentation of sugar to ethyl alcohol — it should be mentioned that sugar is a substance in the waste liquor during sulfite pulp production. He also put into practice a procedure to obtain paper glue from the waste liquor. Both processes were implemented in his plant located in Hof, Germany, in the year 1898 [30].

In 1927 the American Marathon Corporation assigned a group of chemists and engineers to the task of developing commercial products from the organic solids in the spent sulfite liquor from the Marathon's Rothschild pulp and paper operations close to Wausau, Wisconsin, USA. The first products to show promise were leather tanning agents. Later, the characteristics of lignin as dispersing agents became evident. By the mid 1930s, with a considerable amount of basic research accomplished, Marathon transferred operations from a research pilot plant to full-scale production [31].

One of the most well known examples is the production of furfural by the Quaker Oats Company since 1922, thus coupling food, i.e. oat flakes, production and chemical products obtained from the waste [13] (see 2.2). On the basis of furfural a whole section of chemical production developed — furan chemistry.

Agribusiness, especially, strived to achieve combined production from the very beginning. Modern corn refining started in the middle of the 18th century when T. Kingsford commenced operation of his corn refining plant in Oswego, New York [32]. Corn refining is distinguished from corn milling because the refining process separates corn grain into its components, for example starch, fiber, protein and oil, and starch is further processed into a substantial number of products [33].

The extensive usage of green crops has been aim of industry for decades, because there are several advantages. Particularly worthy of mention is the work of Osborn (1920) and Slade and Birkinshaw (1939) on the extraction of proteins from green crops, for example grass or alfalfa [34].

In 1937 N.W. Pirie developed the technical separation and extraction methods needed for this use of green crops [35, 36]. By means of sophisticated methods all the botanical material should have been used, both for production of animal feed, isolated proteins for human nutrition, and as raw material for further industrial processes, for example glue production. The residual material, juices rich in nutrients, had initially been used as fertilizer; later they were used for generation of fermentation heat based on biogas production [37, 38].

These developments resulted in market-leading technology, for example the Proxan and Alfaprox procedures, used for generation of protein–xanthophyll concentrates, including utilization of the by-products, however, predominantly in agriculture [39].

In the United States commercial production of chlorophyll and carotene by extraction from alfalfa leaf meal had started in 1930 [40, 41]. For example Strong, Cobb and Company produced 0.5 t chlorophyll per day from alfalfa as early as 1952. The water-soluble chlorophyll, or chlorophyllin, found use as deodorizing agent in toothpastes, soaps, shampoos, candy, deodorants, and pharmaceuticals [42].

A historical important step for today's biorefinery developments was the industry-politics-approach of "Chemurgy", founded in 1925 in the US by the chemist W.J. Hale, son-in-law of H. Dow, the founder of Dow Chemical, and C. H. Herty, a former President of the American Chemical Society. They soon found prominent support from H. Ford and T. A. Edisons. Chemurgy, an abbreviation of "chemistry" and "ergon", the Greek word for work [43], means by analogy "chemistry from the acre" that is the connection of agriculture with the chemical industry.

Chemurgy was soon shown to have a serious industrial political philosophy — the objective of utilizing agricultural resources, nowadays called renewable resources, in industry. There have been common conferences between agriculture, industry, and science since 1935 with a national council called the "National Farm Chemurgic Council" [44]. The end of Chemurgy started with the flooding of the world market with cheap crude oil after World War II; numerous inventions and production processes remained, however, and are again highly newsworthy. One was a car, introduced by Henry Ford 1941, whose car interior lining and car body consisted 100 % of bio-synthetics; to be specific it had been made from a cellulose meal, soy meal, formaldehyde resin composite material in the

proportions 70 %:20 %:10 %, respectively. The alternative fuel for this car was pyrolysis methanol produced from cannabis. Throughout the thirties more than 30 industrial products based on soy bean were created by researchers from the Ford company; this made it necessary to apply complex conversion methods [45]. Hale was a pioneer of ethyl alcohol and hydrocarbon fuel mixture (power alcohol, gasohol) [46]. This fuel mixture, nowadays called E10-fuel, consisting of 10 % bioethanol and 90 % hydrocarbon-based fuel, has been the national standard since the beginning of this millennium in the United States.

Associated with the work of Bergius, 1933 [47], and Scholler, 1923 and 1935 [48, 49], wood saccharification was reanimated at the end of World War II. Beside optimization of the process, use of lignocelluloses was of great interest. The continuously growing agribusiness left behind millions of tons of unused straw. Two Americans, Othmer and Katzen, were the main pioneers in the field of wood saccharification [50]. Between the years 1935 and 1960 several hydrolysis plants were built in Germany and the United States; in these deal, wood flour, surplus lumber, and also straw were hydrolyzed [51]. One of the most well known plants are those of Scholler/Tornesch located in Tornesch, Germany, with a production rate of 13,000 t/a, in Dessau, Germany, production rate 42,000 t/a, based on wood, in Holzminden, Germany, production rate of 24,000 t/a, also based on wood, in Ems, Switzerland, with a production rate of 35,000 t/a, also wood based, and the plants in Madison and Springfield, United States, and the Bergius plants in Rheinau, Germany (Rheinau I, built 1930, with a production rate of 8000 t/a, based on surplus lumber; Rheinau II, built 1960 with a production rate of 1200 t/a, based on wood) and the plant in Regensburg, Germany, with a production rate of 36,000 t/a [52].

During World War II the plant in Springfield, Oregon, US (using the Scholler-Tornesch process as modified by Katzen) produced 15,000 gallons of ethyl alcohol per day from 300 t wood flour and sawdust, i.e. 50 gallons per ton of wood [53]. The plant in Tornesch, Germany, has been producing approximately 200 liters of ethyl alcohol, purity 100 %, per ton wood and approximately 40 kg yeast per ton of wood. In 1965 there were 14 plants in the former Soviet Union, with a total capacity of 700,000 t/a and an overall annual wood consumption of 4×10^6 t [9].

During the nineteen-sixties wood chemistry had its climax. Projects had been developed, which made it possible to produce nearly all chemical products on the basis of wood. Examples are the complex chemical technological approaches of wood processing from Timell 1961 [54], Stamm 1964 [55], James 1969 [56], Brink and Pohlmann 1972 [57], and the wood-based chemical product trees by Oshima 1965 [58]. Although these developments did not make their way into industrial production, they are an outstanding platform for today's lignocellulose conversions, product family trees, and LCF biorefineries (5.2).

Most of the above mentioned technologies and products, some of which were excellent, could not compete with the fossil-based industry and economy; nowadays, however, they are prevailing again. The basis for this revival started in the seventies, when the oil crisis and continuously increasing environmental pollution resulted in a broad awareness that plants could be more than food and animal feed. At the same time the disadvantages of intensive agricultural usage, for example over-fertilization, soil erosion, and the enormous amounts of waste, were revealed. From this situation developed complex concepts, which have been published, in which the aim was, and still is, technological and economical cooperation of agriculture, forestry, the food-production industry, and conventional industry, or at least consideration of integrated utilization of renewable resources.

Typical examples of this thinking were:

- integrated industrial utilization of wood and straw [59];
- industrial utilization of fast growing wood-grass [60, 61];
- complex utilization of green biomass, for example grass and alfalfa, by agriculture and industry [62 – 64];
- corn wet-grinding procedures with associated biotechnological and chemical product lines [65];
- modern aspects of thermochemical biomass conversion [66];
- discussion of the concept "organic chemicals from biomass" with main focus on biotechnological methods and products (white biotech-

nology) [67 – 69] and industrial utilization of biomass [70].

These rich experiences of the industrial utilization of renewable resources, new agricultural technology, biotechnology, and chemistry, and the changes in ecology, economics, and society led inevitably to the topic of complex and integrated substantial and energy utilization of biomass and, finally, to the biorefinery.

3. Situation

3.1. Some Current Aspects of Biorefinery Research and Development

Since the beginning of the 1990s the utilization of renewable resources for production of non-food products has fostered research and development which has received increasing attention from industry and politicians [71 – 73]. Integrated processes, biomass refinery technology, and biorefinery technology have become object of research and development. Accordingly, the term "biorefinery" was established in the 1990s [4, 74 – 83]. The respective biorefinery projects are focused on the fabrication of fuels, solvents, chemicals, plastics, and food for human beings. In some countries these biorefinery products are made from waste biomass. At first the main processes in the biorefinery involved ethanol fermentation for fuels (ethanol-oriented biorefineries) [84 – 88], lactic acid (LA) fermentation [28, 89], propanediol (PDO) fermentation [90], and the lysine fermentation [91] especially for polymer production. The biobased polymers poly(lactic acid) [28], propandiol-derived polymers [92], and polylysine [91] have been completed by polyhydroxyalkanoates [93] and polymerized oils [94].

Many hybrid technologies were developed from different fields, for example bioengineering, polymer chemistry, food science and agriculture. Biorefinery systems based on cereals [95, 96], lignocelluloses [97, 98], and grass and alfalfa [38, 99], and biorefinery optimization tools are currently being developed [100, 101]. The integration of molecular plant genetics to support the raw material supply is currently being discussed intensely [102, 103].

Broin and Associates has begun the development of a second generation of dry mill refineries and E.I. du Pont de Nemours has developed an integrated corn-based biorefinery. In 2001 NatureWorks LCC (to Cargill, former Cargill Dow LLC) started the industrial production of PLA (PLA-oriented Biorefinery) on the basis of maize.

Biorefineries are of interest ecologically [104], economically, and to business, government, and politicians [105 – 110]. National programs [111, 112], biobased visions [113], and plans [114] have been developed and the international exchange of information is increasing, for example as a result, among others, of series of international congresses and symposia:

- BIO World Congress on Industrial Biotechnology and Bioprocessing [115];
- biomass conferences [116, 117];
- The Green and Sustainable Chemistry Congress [118];
- ACS Symposium Series and [119]
- the Biorefinica symposia series [120, 121].

Currently, biorefinery systems are in the stage of development world-wide. An overview of the main aspects, activity, and discussions is the content of this book. An attempt to systematize the topic "Biorefinery" will be presented below.

3.2. Raw Material Biomass

Nature is a permanently renewing production chain for chemicals, materials, fuels, cosmetics, and pharmaceuticals. Many of the biobased industry products currently used are results of direct physical or chemical treatment and processing of biomass, for example cellulose, starch, oil, protein, lignin, and terpenes. On one hand one must mention that because of the help of biotechnological processes and methods, feedstock chemicals are produced such as ethanol, butanol, acetone, lactic acid, and itaconic acid, as also are amino acids, e.g., glutaminic acid, lysine, tryptophan. On the other hand, only 6×10^9 t of the yearly produced biomass, $1.7-2.0 \times 10^{11}$ t, are currently used, and only 3 to 3.5 % of this amount is used in non-food applications, for example chemistry [122].

The basic reaction of biomass is photosynthesis according to:

Industrial utilization of raw materials from agriculture, forestry, and agriculture is only just beginning. There are several definitions of the term "biomass" [122]:

- the complete living, organic matter in our ecological system (volume/non-specific);
- the plant material constantly produced by photosynthesis with an annual growth of 170 billion tons (marine plants excluded);
- the cell-mass of plants, animals, and microorganism used as raw materials in microbiological processes.

Biomass is defined in a recent US program [111, 112]: "The term "biomass" means any organic matter that is available on a renewable or recurring basis (excluding old-growth timber), including dedicated energy crops and trees, agricultural food and feed crop residues, aquatic plants, wood and wood residues, animal wastes, and other waste materials."

For this reason it is essential to define biomass in the context of the industrial utilization. A suggestion for a definition of "industrial biomass" [111, 112] is: "The term "industrial biomass" means any organic matter that is available on a renewable or recurring basis (excluding old-growth timber), including dedicated energy crops and trees, agricultural food and feed crop residues, aquatic plants, wood and wood residues, animal wastes, and other waste materials usable for industrial purposes (energy, fuels, chemicals, materials) and include wastes and co-wastes of food and feed processing."

Most biological raw material is produced in agriculture and forestry and by microbial systems. Forestry plants are excellent raw materials for the paper and cardboard, construction, and chemical industries. Field fruits are a pool of organic chemicals from which fuels, chemicals, chemical products, and biomaterials are produced (Fig. 2), [72].

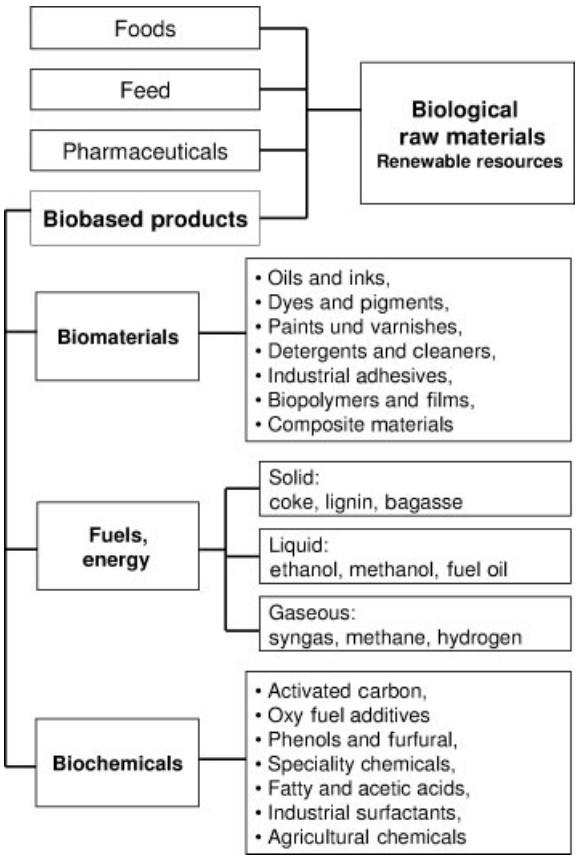

Figure 2. Products and product classes based on biological raw materials [81]

Advances in genetics, biotechnology, process chemistry, and engineering are leading to a new manufacturing concept for converting renewable biomass to valuable fuels and products, generally referred to as the biorefinery. The integration of agroenergy crops and biorefinery manufacturing technologies offers the potential for the development of sustainable biopower and biomaterials that will lead to a new manufacturing paradigm [2, 123].

Waste biomass and biomass of nature and agricultural cultivation are valuable organic reservoirs of raw material and must be used in accordance with their organic composition. During the development of biorefinery systems the term "waste biomass" will become obsolete in the medium-term [124].

3.3. National Vision and Goals and Plan for Biomass Technology in the United States

Industrial development was pushed by the US President [111] and by the US Congress [112], initially in 2000. In the USA it is intended that by 2020 at least 25 % of organic-carbon-based industrial feedstock chemicals and 10 % of liquid fuels (compared with levels in 1994) will be produced by biobased industry. This would mean that more than 90 % of the consumption of organic chemicals in the US and up to 50 % of liquid fuel needs would be biobased products [4].

The Biomass Technical Advisory Committee (BTAC) of the USA in which leading representatives of industrial companies, for example Dow Chemical, E.I. du Pont de Nemours, Cargill Dow LLC, and Genencor International, and Corn growers associations and the Natural Resources Defense Council are involved and which acts as advisor to the US government, has made a detailed plan with steps toward targets of 2030 with regard to bioenergy, biofuels, and bioproducts (Table 1) [113].

Simultaneously, the plan *Biomass Technology in the United States* has been published [114] in which research, development, and construction of biorefinery demonstration plants are determined. Research and development are necessary to:

1) increase scientific understanding of biomass resources and improve the tailoring of those resources;
2) improve sustainable systems to develop, harvest, and process biomass resources;
3) improve efficiency and performance in conversion and distribution processes and technologies for development of a host of biobased products and
4) create the regulatory and market environment necessary for increased development and use of biobased products.

The Biomass Advisory Committee has established specific research and development objectives for feedstock production research. Target crops should include oil and cellulose-producing crops that can provide optimum energy content and usable plant components. Currently, however, there is a lack of understanding of plant biochemistry and inadequate genomic and metabolic information about many potential crops. Specific research to produce enhanced enzymes and chemical catalysts could advance biotechnology capabilities.

3.4. Vision and Goals and Plan for Biomass Technology in the European Union and Germany

In Europe there are already regulations about substitution of nonrenewable resources by biomass in the field of biofuels for transportation [125] and the "Renewable energy law" of 2000 [126]. According to the EC Directive "On the promotion of the use of biofuels" the following products are regarded as "biofuels":

(a) "bioethanol", (b) "biodiesel", (c) "biogas", (d) "biomethanol", (e) "biodimethyl ether",

(f) "bio-ETBE (ethyl tertiary-butyl ether)" on the basis of bioethanol,

(g) "bio-MTBE (methyl tertiary butyl ether)" on the basis of biomethanol, and

(h) "synthetic biofuels", (i) "biohydrogen", (j) pure vegetable oil

Member States of the EU have been requested to define national guidelines for a minimum amounts of biofuels and other renewable fuels (with a reference value of 2 % by 2005 and 5.75 % by 2010 calculated on the basis of energy

Table 1. The US national vision goals for biomass technologies by the Biomass Technical Advisory Committee [113]

Year	2002	2010	2020	2030
BioPower (BioEnergy)Biomass share of electricity and heat demand in utilities and industry	2.8 % (2.7 quad)[a]	4 % (3.2 quad)	5 % (4.0 quad)	5 % (5.0 quad)
BioFuels Biomass share of demand for transportation fuels	0.5 % (0.15 quad)	4 % (1.3 quad)	10 % (4.0 quad)	20 % (9.5 quad)
BioProducts Share of target chemicals that are biobased	5 %	12 %	18 %	25 %

[a] 1 quad=1 quadrillion BTU=1 German billiarde BTU; BTU=British thermal unit; 1 BTU=0.252 kcal, 1 kW=3413 BTU, 1 kcal=4.186 kJ

content of all petrol and diesel fuels for transport purposes). Table 2 summarizes this goal of the EU and also those of Germany with regard to establishment of renewable energy and biofuel [127, 128].

Today there are no guidelines concerning "biobased products" in the European Union and in Germany. After passing directives relating to bioenergy and biofuels, however, such a decision is on the political agenda. The "biofuels" directive already includes ethanol, methanol, dimethyl ether, hydrogen, and biomass pyrolysis which are fundamental product lines of the future biobased chemical industry.

In the year 2003, an initiative group called "Biobased Industrial Products" consisting of members from industry, small and middle-class businesses, and research, and development facilities met and formulated a strategy paper, called "BioVision 2030" [129]. This strategy paper has been included in the resolution of the German Government (Deutscher Bundestag) on the topic "Accomplish basic conditions for the industrial utilization of renewable resources in Germany" [130]. An advisory committee consisting of members of the chemical industry, related organizations, research facilities, and universities has been established to generate a plan concerning the formulation of the objectives for the third column, bio-products in Europe (Table 2) [131].

A recent vision paper published by the Industrial Biotechnology Section of the European Technology platform for Sustainable Chemistry foresee up to 30 % of raw materials for chemical industry coming from renewable sources by 2025 [132]. Lately European Commission and U S Department of Energy are come to an agreement for cooperation on this field [133].

4. Principles of Biorefineries

4.1. Fundamentals

Biomass, similar to petroleum, has a complex composition. Its primary separation into main groups of substances is appropriate. Subsequent treatment and processing of those substances lead to a whole range of products. Petrochemistry is based on the principle of generating simple to handle and well defined chemically pure products from hydrocarbons in refineries. In efficient product lines, a system based on family trees has been built, in which basic chemicals, intermediate products, and sophisticated products are produced. This principle of petroleum refineries must be transferred to biorefineries. Biomass contains the synthesis performance of the nature and has different C:H:O:N ratio from petroleum. Biotechnological conversion will become, with chemical conversion, a big player in the future (Fig. 3).

Thus biomass can already be modified within the process of genesis in such a way that it is adapted to the purpose of subsequent processing, and particular target products have already been formed. For those products the term "precursors" is used. Plant biomass always consists of the basic products carbohydrates, lignin, proteins, and fats, and a variety of substances such as vitamins, dyes, flavors, aromatic essences of very different chemical structure. Biorefineries combine the essential technologies which convert biological raw materials into the industrial intermediates and final products (Fig. 4).

A technically feasible separation operation, which would enable separate use or subsequent processing of all these basic compounds, is currently in its initial stages only. Assuming that of

Table 2. Targets of the EU and Germany with regard to the introduction of technologies based on renewable resources

Year	2001	2005	2010	2020–2050
Bioenergy Share of wind power, photovoltaics, biomass and geothermal electricity and heat demand in utilities and industry	7.5 %	–	12.5 %	26 % (2030)58 % (2050)
Biofuels Biomass share of demand in transportation fuels (petrol and diesel fuels)	1.4 %	2.8 %	5.75 %	20 % (2020)
Biobased Products Share of target chemicals that are biobased	8 %			

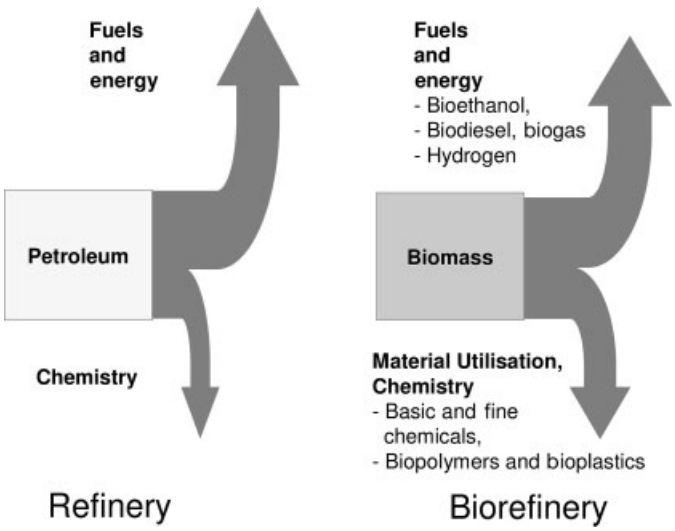

Figure 3. Comparison of the basic-principles of the petroleum refinery and the biorefinery

the estimated annual production of biomass by biosynthesis of 170 billion tons 75 % is carbohydrates, mainly in the form of cellulose, starch, and saccharose, 20 % lignin, and only 5 % other natural compounds such as fats (oils), proteins, and other substances [134], the main attention should first be focused on efficient access to carbohydrates, and their subsequent conversion to chemical bulk products and corresponding final products. Glucose, accessible by microbial or chemical methods from starch, sugar, or cellulose, is, among other things, predestined for a key position as a basic chemical, because a broad range of biotechnological or chemical products is accessible from glucose [135]. For starch the advantage of enzymatic compared with chemical hydrolysis is already known [136].

For cellulose this is not yet realized. Cellulose-hydrolyzing enzymes can only act effectively after pretreatment to break up the very stable lignin/cellulose/hemicellulose composites [137]. These treatments are still mostly thermal, thermomechanical, or thermochemical, and require considerable input of energy. The arsenal for microbial conversion of substances from glucose is large, and the reactions are energetically profitable. It is necessary to combine degradation processes via glucose to bulk chemicals with the building processes to their subsequent products and materials (Fig. 5).

Among the variety of microbial and chemical products possibly accessible from glucose, lactic acid, ethanol, acetic acid, and levulinic acid, in particular, are favorable intermediates for generation of industrially relevant product family trees. Here, two potential strategies are considered: first, development of new, possibly biologically degradable products (follow-up products from lactic and levulinic acids) or, second, entry as intermediates into conventional product lines

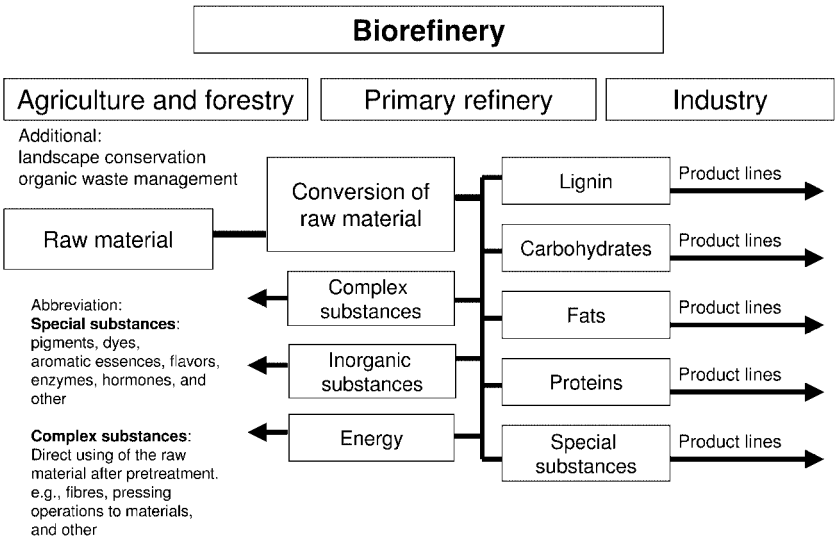

Figure 4. Providing code-defined basic substances (via fractionation) for development of relevant industrial product family trees [81, 82]

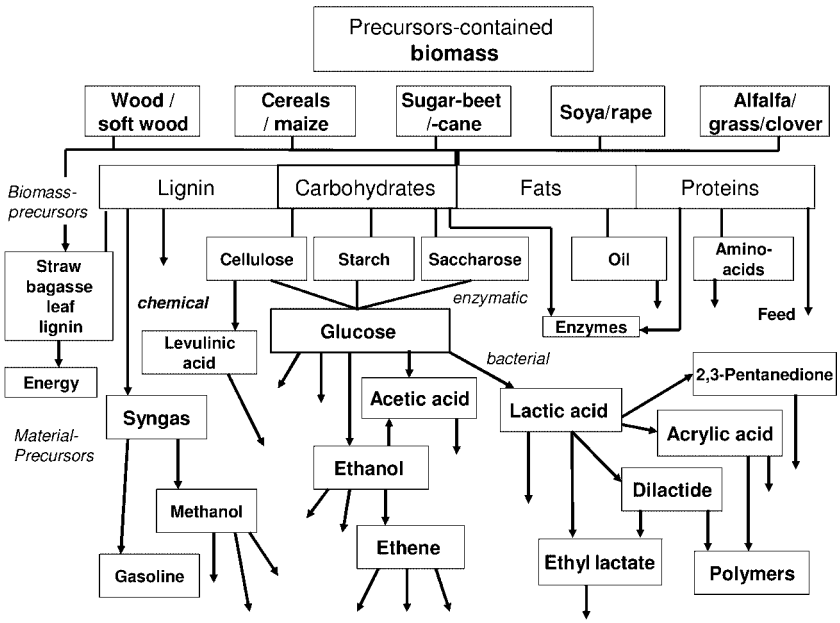

Figure 5. Possible schematic diagram of biorefinery for precursor-containing biomass with preference for carbohydrates [81, 82]

(acrylic acid, 2,3-pentanedione) of petrochemical refineries [81].

4.2. Definition of the Term "Biorefinery"

The young working field "Biorefinery Systems" in combination with "Biobased Industrial Products" is, in various respects, still an open field of knowledge. This is also reflected in the search for an appropriate description. A selection is given below.

The term "Green Biorefinery" was been defined in the year 1997 as: "Green biorefineries represent complex (to fully integrated) systems

of sustainable, environmentally and resource-friendly technologies for the comprehensive (holistic) material and energetic utilization as well as exploitation of biological raw materials in form of green and residue biomass from a targeted sustainable regional land utilization" [76]. The original term used in Germany "complex construction and systems" was substituted by "fully integrated systems". The US Department of Energy (DOE) uses the following definition [138]: "A biorefinery is an overall concept of a processing plant where biomass feedstocks are converted and extracted into a spectrum of valuable products. Based on the petrol-chemical refinery." The American National Renewable Energy Laboratory (NREL) published the definition [139]: "A biorefinery is a facility that integrates biomass conversion processes and equipment to produce fuels, power, and chemicals from biomass. The biorefinery concept is analogous to today's petroleum refineries, which produce multiple fuels and products from petroleum. Industrial biorefineries have been identified as the most promising route to the creation of a new domestic biobased industry."

A general definition is: "Biorefining is the transfer of the efficiency and logic of fossil-based chemistry and substantial converting industry as well as energy production onto the biomass industry".

There is an agreement about the *objective*, which is briefly defined as: "Developed biorefineries, so called "phase III-biorefineries" or "generation III-biorefineries", start with a biomass–feedstock-mix to produce a multiplicity of most various products by a technologies-mix" [77] (Fig. 6).

An example of the type "generation-I biorefinery" is a dry milling ethanol plant. It uses grain as a feedstock, has a fixed processing capability, and produces a fixed amount of ethanol, feed co-products, and carbon dioxide. It has almost no flexibility in processing. Therefore, this type can be used for comparable purposes only.

An example of a type "generation-II biorefinery" is the current wet milling technology. This technology uses grain feedstock, yet has the capability to produce a variety of end products depending on product demand. Such products include starch, high-fructose corn syrup, ethanol, corn oil, plus corn gluten feed, and meal. This type opens numerous possibilities to connect industrial product lines with existing agricultural production units. "Generation-II biorefineries" are, furthermore, plants like NatureWorks PLA facility [1, 28] (see Sections 2 and 3.1) or ethanol biorefineries, for example Iogen's wheat straw to ethanol plant [147].

Third generation (generation-III) and more advanced biorefineries have not yet been built but will use agricultural or forest biomass to produce multiple products streams, for example ethanol for fuels, chemicals, and plastics [140, 147].

4.3. The Role of Biotechnology

The application of biotechnological methods will be highly important with the development of biorefineries for production of basic chemicals, intermediate chemicals, and polymers [141 – 143]. The integration of biotechnological methods must be managed intelligently in respect of physical and chemical conversion of the biomass. Therefore the biotechnology cannot remain limited to glucose from sugar plants and starch from starch-producing plants. One main objective is the economic use of biomass containing lignocellulose and provision of glucose in the family tree system. Glucose is a key chemical for microbial processes. The preparation of a large number of family tree-capable basic chemicals is shown in subsequent sections.

4.3.1. Guidelines of Fermentation Section within Glucose-product Family Tree

Among the variety of chemical products, and derivatives of these, accessible microbially from glucose a product family tree can be developed, for example (C-1)-chemicals methane, carbon dioxide, methanol; (C-2)-chemicals ethanol, acetic acid, acetaldehyde, ethylene, (C-3)-chemicals lactic acid, propanediol, propylene, propylene oxide, acetone, acrylic acid, (C-4)-chemicals diethyl ether, acetic acid anhydride, malic acid, vinyl acetate, *n*-butanol, crotonaldehyde, butadiene, 2,3-butanediol, (C-5)-chemicals itaconic acid, 2,3-pentane dione, ethyl lactate, (C-6)-chemicals sorbic acid, parasorbic acid, citric acid, aconitic acid, isoascorbinic

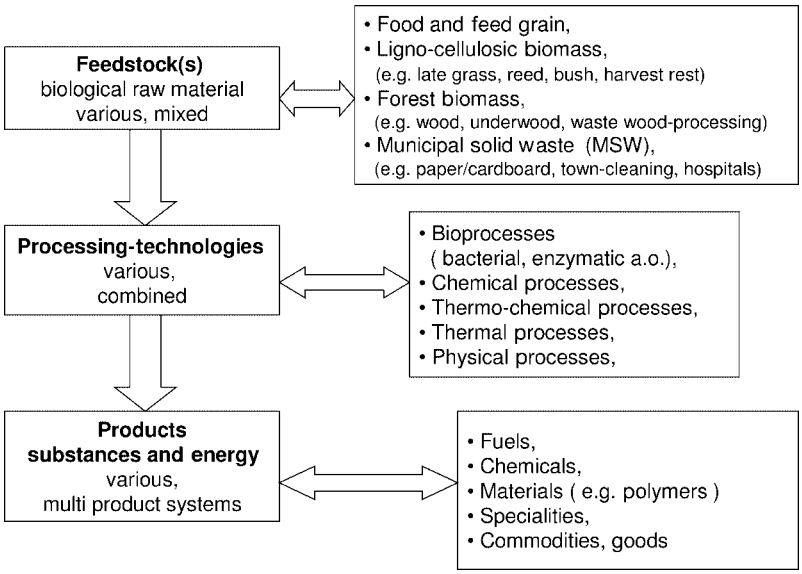

Figure 6. Basic principles of a biorefinery (generation III biorefinery) [81]

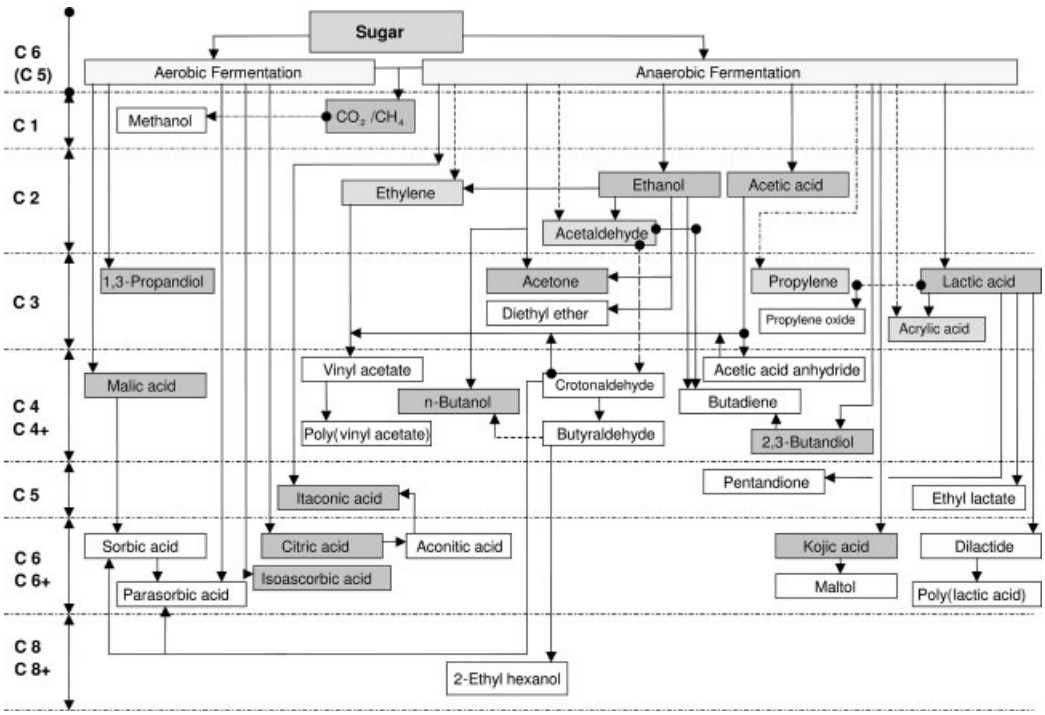

Figure 7. Biotechnological sugar-based product family tree

acid, kojic acid, maltol, dilactide, (C-8)-chemicals 2-ethyl hexanol (Fig. 7).

Guidelines are currently being developed for the fermentation section of a biorefinery. The question of efficient arrangement of the technological design for production of bulk chemicals needs an answer. Considering the manufacture of lactic acid and ethanol, the basic technological operations are very similar. Selection of biotechnologically based products from biore-

fineries should be done in a way such that they can be produced from the substrates glucose or pentoses. Furthermore the fermentation products should be extracellular. Fermenters should have batch, feed batch, or CSTR design. Preliminary product recovery should require steps like filtration, distillation, or extraction. Final product recovery and purification steps should possibly be product-unique. In addition, biochemical and chemical processing steps should be advantageously connected.

Unresolved questions for the fermentation facility include:

1) whether or not the entire fermentation facility can/should be able to change from one product to another;
2) whether multiple products can be run in parallel, with shared use of common unit operations;
3) how to manage scheduling of unit operations; and
4) how to minimize in-plant inventories, while accommodating necessary change-over between different products in the same piece of equipment [98].

4.4. Building Blocks, Chemicals and Potential Screening

A team from Pacific Northwest National Laboratory (PNNL) and NREL submitted a list of twelve potential biobased chemicals [101]. A key area of the investigation as biomass precursors, platforms, building blocks, secondary chemicals, intermediates, products and uses (Fig. 8).

The final selection of 12 building blocks began with a list of more than 300 candidates. A shorter list of 30 potential candidates was selected by using an iterative review process based on the petrochemical model of building blocks, chemical data, known market data, properties, performance of the potential candidates, and previous industry experience of the team at PNNL and NREL. This list of 30 was ultimately reduced to 12 by examining the potential markets for the building blocks and their derivatives and the technical complexity of the synthetic pathways.

The reported block chemicals can be produced from sugar by biological and chemical conversions. The building blocks can be subsequently converted to several high-value biobased chemicals or materials. Building-block chemicals, as considered for this analysis, are molecules with multiple functional groups with the potential to be transformed into new families of useful molecules. The twelve sugar-based building blocks are 1,4-diacids (succinic, fumaric, and malic), 2,5-furandicarboxylic acid, 3-hydroxypropionic acid, aspartic acid, glucaric acid, glutamic acid, itaconic acid, levulinic acid, 3-hydroxybutyrolactone, glycerol, sorbitol, and xylitol/arabinitol [101].

A second-tier group of building blocks was also identified as viable candidates. These include gluconic acid, lactic acid, malonic acid, propionic acid, the triacids citric and aconitic, xylonic acid, acetoin, furfural, levuglucosan, lysine, serine, and threonine. Recommendations for moving forward include:

- examining top value products from biomass components, for example aromatic compounds, polysaccharides, and oils;
- evaluating technical challenges in more detail in relation to chemical and biological conversion; and
- increasing the suites of potential pathways to these candidates.

The co-production of organic acids, ethanol, and electricity, as well as the opportunities for their integration, will improve overall economics. [132]. Also, the possibilities of catalysis for the production of platform chemicals have not yet sufficiently been exploited [144]. In addition, further investigations on microbial biology and bioprocessing are anticipated [145].

No further down select products from syngas was undertaken. For the purposes of the study from PNNL and NREL hydrogen and methanol comprise the best short-term prospects for biobased commodity chemical production, because obtaining simple alcohols, aldehydes, mixed alcohols, and Fischer-Tropsch liquids from biomass is not economically viable and requires additional development [101].

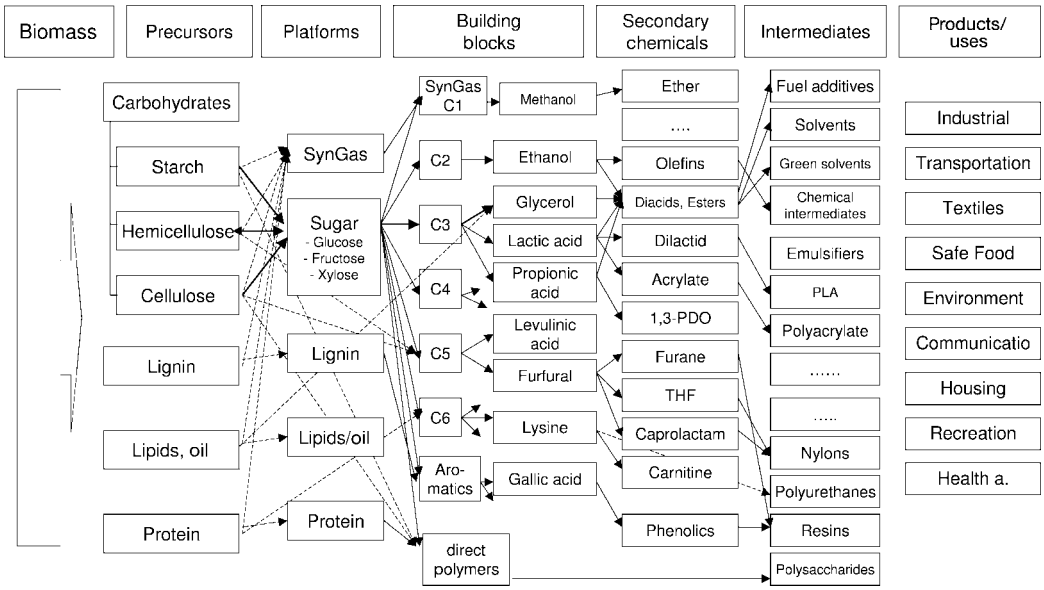

Figure 8. Model of a biobased product flow-chart for biomass feedstock [101]

5. Biorefinery Systems and Design

5.1. Introduction

Biobased products are prepared for a usable economic use by meaningful combination of different methods and processes (physical, chemical, biological, and thermal). It is therefore necessary that basic biorefinery technologies are developed. For this reason profound interdisciplinary cooperation of various disciplines in research and development is inevitable. It seems reasonable, therefore, to refer to the term "biorefinery design", which means: "bringing together well founded scientific and technological basics, with similar technologies, products, and product lines, inside biorefineries". The basic conversions of each biorefinery can be summarized as follows. In the first step, the precursor-containing biomass is separated by physical methods. The main products (M_1–M_n) and the by-products (B_1–B_n) will subsequently be subjected to microbiological or chemical methods. The follow-up products (F_1–F_n) of the main and by-products can also be converted or enter the conventional refinery (Fig. 6).

Currently four complex biorefinery systems are used in research and development:

1) the "lignocellulosic feedstock biorefinery" which use "nature-dry" raw material, for example cellulose-containing biomass and waste;
2) the "whole crop biorefinery" which uses raw material such as cereals or maize;
3) the "green biorefineries" which use "nature-wet" biomass such as green grass, alfalfa, clover, or immature cereal [81, 82]; and
4) the "biorefinery two platforms concept" includes the sugar platform and the syngas platform [101].

5.2. Lignocellulosic Feedstock Biorefinery

Among the potential large-scale industrial biorefineries the lignocellulose feedstock (LCF) biorefinery will most probably be pushed through with the greatest success. On the one side the raw material situation is optimum (straw, reed, grass, wood, paper-waste, etc.), on the other side conversion products have a good position on both the traditional petrochemical and future biobased product market [146, 148, 149]. An important point for utilization of biomass as chemical raw material is the cost of raw material. Currently the cost of corn stover or straw is 30 US$/ton and that of corn is 110 US$/ton (3 US$/ bushel; US bushel corn =25.4012 kg=56 lb) [150].

Lignocellulose materials consist of three primary chemical fractions or precursors:

- hemicellulose/polyoses, sugar polymers of, predominantly, pentoses;
- cellulose, a glucose polymer; and
- lignin, a polymer of phenols (Fig. 9).

$$\text{Lignocellulose} + H_2O \rightarrow \text{Lignin} + \text{Cellulose} + \text{Hemicellulose}$$
$$\text{Hemicellulose} + H_2O \rightarrow \text{Xylose}$$
$$\text{Xylose } (C_5H_{10}O_5) + \text{acid Catalyst} \rightarrow \text{Furfural } (C_5H_4O_2) + 3H_2O$$
$$\text{Cellulose}(C_6H_{10}O_5) + H_2O \rightarrow \text{Glucose } (C_6H_{12}O_6)$$

Figure 9. A possible general equation for conversion at the LCF biorefinery

The lignocellulosic biorefinery-regime is distinctly suitable for genealogical compound trees. The main advantages of this method is that the natural structures and structural elements are preserved, the raw materials are inexpensive, and large product varieties are possible (Fig. 10). Nevertheless there is still a demand for development and optimization of these technologies, e.g., in the field of separation of cellulose, hemicellulose and lignin, and utilization of the lignin in the chemical industry.

An overview of potential products of an LCF biorefinery is shown in Figure 11. In particular furfural and hydroxymethylfurfural are interesting products. Furfural is a starting material for production of nylon 6,6 and nylon 6. The original process for production of nylon-6,6 was based on furfural (see also 2.2). The last of these production plants was closed in 1961 in the USA, for economic reasons (the artificially low price of petroleum). Nevertheless the market for nylon 6 is huge. Catalytic transformation of these future feedstocks provides new market opportunities. Levulinic acid is one of 12 bioderived feedstocks and it can be obtained directly by cellulose hydrolysis [101]. A number of high value derivatives, such as pyrrolidones, monomers for the synthesis of high Tg amorphous polymers and fuel additives, such as ethyl levulinate can be prepared [123, 151].

There are, however, still some unsatisfactory aspects of the LCF, for example the utilization of lignin as fuel, adhesive, or binder. Unsatisfactory because the lignin scaffold contains substantial amounts of mono-aromatic hydrocarbons, which, if isolated in an economically efficient way, could add a significant increase in value to the primary processes. It should be noticed there are no natural enzymes capable of splitting the naturally formed lignin into basic monomers as easily as is possible for natural polymeric carbohydrates or proteins [152].

An attractive process accompanying the biomass-nylon-process is the already mentioned hydrolysis of the cellulose to glucose and the production of ethanol. Some yeasts cause disproportionation of the glucose molecule during their generation of ethanol from glucose, which shifts almost all its metabolism into ethanol production, making the compound obtainable in 90 % yield (w/w; with regard to the chemical equation for the process).

On the basis of recent technology a plant has been conceived for production of the main products furfural and ethanol from LC feedstock from West Central Missouri (USA). Optimal

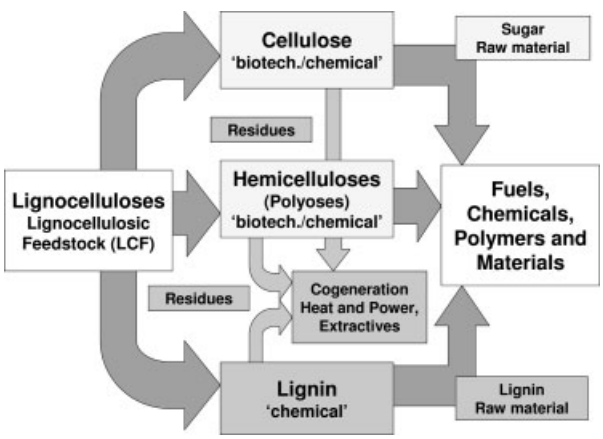

Figure 10. Lignocellulosic feedstock biorefinery

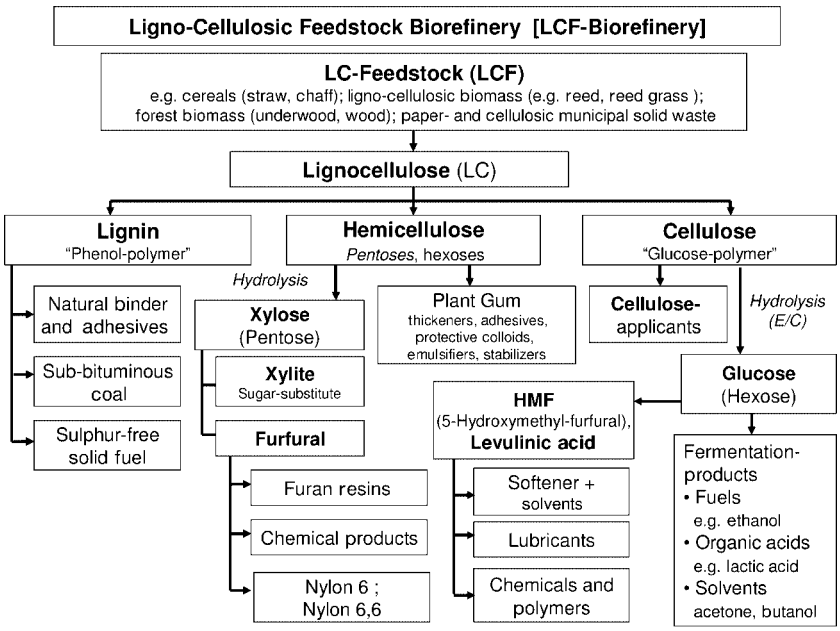

Figure 11. Products of a lignocellulosic feedstock biorefinery (LCF-biorefinery, Phase III) [81, 82, 98]

profitability can be achieved with a daily consumption of approximately 4360 tons of feedstock. The plant produces 47.5 million gallon of ethanol and 323 tons of furfural annually [77].

Ethanol can be used as a fuel additive. It is also a connecting product to the petrochemical refinery, because it can be converted into ethene by chemical methods and it is well-known that ethene is at the start of a series of large-scale technical chemical syntheses for production of important commodities such as polyethylene or poly(vinyl acetate). Other petrochemically produced substances, for example hydrogen, methane, propanol, acetone, butanol, butandiol, itaconic acid, and succinic acid, can also be manufactured by microbial conversion of glucose [153, 154].

5.3. Whole-crop Biorefinery

Raw materials for the "whole crop biorefinery" are cereals such as rye, wheat, triticale, and maize. The first step is mechanical separation into corn and straw, approximately in a 1:1.1–1.3 ratio, respectively, see Fig. 12 [155]. Straw is a mixture of chaff, nodes, ears, and leaves. The straw is an LC feedstock and may further be processed in a LCF biorefinery.

There is the possibility of separation into cellulose, hemicellulose, and lignin and their further conversion in separate product lines which are shown in the LCF biorefinery. The straw is also a starting material for production of syngas by pyrolysis technology. Syngas is the basic material for synthesis of fuels and methanol (Fig. 13).

The corn may either be converted into starch or used directly after grinding to meal. Improved molecular disassembly and depolymerization of grain starch to glucose are key to reducing energy use in the bioconversion of glucose to chemicals, ingredients, and fuels. [156, 157]

Further processing may be conducted by four processes — breaking up, plasticization, chemical modification, or biotechnological conversion via glucose. The meal can be treated and finished by extrusion into binder, adhesives, and filler. Starch can be finished by plasticization (co- and mix-polymerization, compounding with other polymers), chemical modification (etherification into carboxymethyl starch; esterification and re-esterification into fatty acid esters via acetic starch; splitting reductive amination into ethylenediamine, etc., hydrogenative splitting into sorbitol, ethylene glycol, propylene glycol, and glycerin), and biotechnological

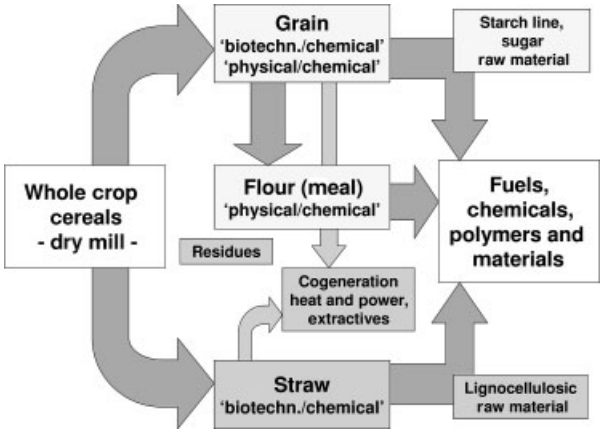

Figure 12. Whole-crop biorefinery based on dry milling

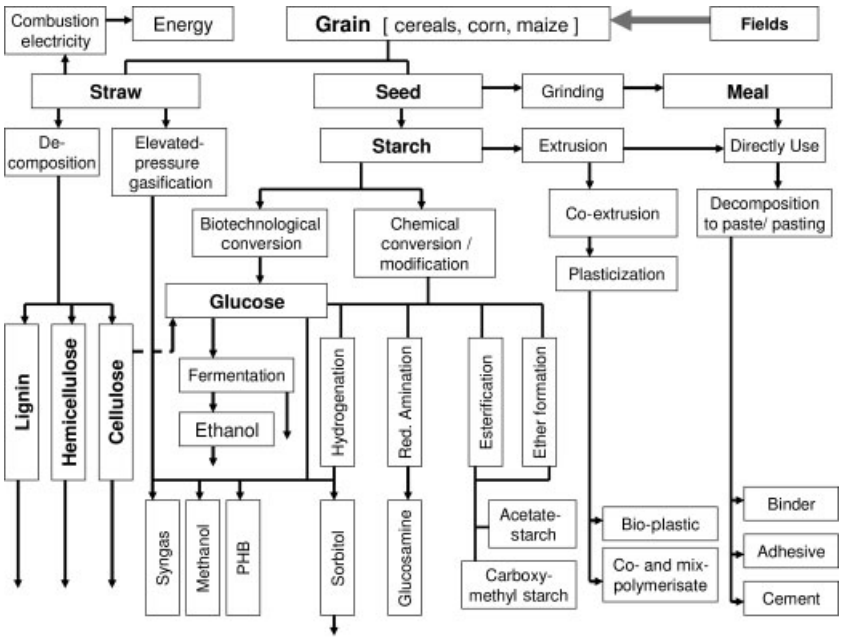

Figure 13. Products from a whole-crop biorefinery [81, 82]

conversion into poly-3-hydroxybutyric acid [72, 79, 95, 96, 158, 159]

An alternative to traditional dry fractionation of mature cereals into grains and straw only was been developed by Kockums Construction (Sweden), which later became Scandinavian Farming. In this crop-harvest system whole immature cereal plants are harvested. The whole harvested biomass is conserved or dried for long-term storage. When convenient, it can be processed and fractionated into kernels, straw chips of internodes, and straw meal (leaves, ears, chaff, and nodes) (see also 5.4).

Fractions are suitable as raw materials for the starch polymer industry, the feed industry, the cellulose industry, and particle board producers, gluten can be used by the chemical industry and as a solid fuel. Such dry fractionation of the whole crop to optimize the utilization of all botanical components of the biomass has been described [160, 161]. A biorefinery and its profitability has been described elsewhere [162].

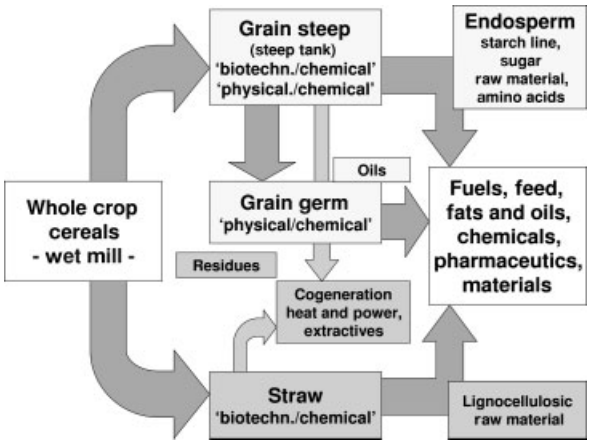

Figure 14. Whole-crop biorefinery, wet-milling

One expansion of the product lines in grain processing is the "whole crop wet mill-based biorefinery". The grain is swelled and the grain germ is pressed, releasing high-value oils. The advantages of whole-crop biorefinery based on wet milling are that production of natural structures and structure elements such as starch, cellulose, oil, and amino acids (proteins) are kept high yet well known basic technology and processing lines can still be used. High raw material costs and, for industrial utilization, the necessary costly source technology are the disadvantages. Some of the products formed command high prices in, e.g., the pharmaceutical and cosmetics industries (Figs. 14 and 15). The basic biorefinery technology of corn wet mills used 11 % of the US corn harvest in 1992, made products worth $7.0 billion, and employed almost 10,000 people [4].

Wet milling of corn yields corn oil, corn fiber, and corn starch. The starch products of the US corn wet milling industry are fuel alcohol (31 %), high-fructose corn syrup (36 %), starch (16 %), and dextrose (17 %). Corn wet milling also generates other products (e.g. gluten meal, gluten feed, oil) [65]. An overview about the product range is shown in Figure 15.

5.4. Green Biorefinery

Green biorefineries are also multi-product systems and furnish cuts, fractions, and products in accordance with the physiology of the corresponding plant material, which maintains and utilizes the diversity of syntheses achieved by nature. Most green biomass is green crops, for example grass from cultivation of permanent grass land, closure fields, nature reserves, or green crops, such as lucerne, clover, and immature cereals from extensive land cultivation. Green crops are a natural chemical factory and food plant and are primarily used as forage and as a source of leafy vegetables. A process called wet-fractionation of green biomass, green crop fractionation, can be used for simultaneous manufacture of both food and non food items [163].

Scientists in several countries, in Europe and elsewhere, have developed green crop fractionation [164 – 166]. Green crop fractionation is now studied in approximately 80 countries [167]. Several hundred temperate and tropical plant species have been investigated for green crop fractionation [166, 168, 169] and more than 300,000 higher plants species have still to be investigated. The subject has been covered by several reviews [76, 164 – 166, 169 – 173]. Green biorefineries can, by fractionation of green plants, process from a few tonnes of green crops per hour (farm scale process) to more than 100 tonnes per hour (industrial scale commercial process).

Careful wet fractionation technology is used as a first step (primary refinery) to isolate the contents of the green crop (or humid organic waste goods) in their natural form. Thus, they are separated into a fiber-rich press cake (PC) and a nutrient-rich green juice (GJ).

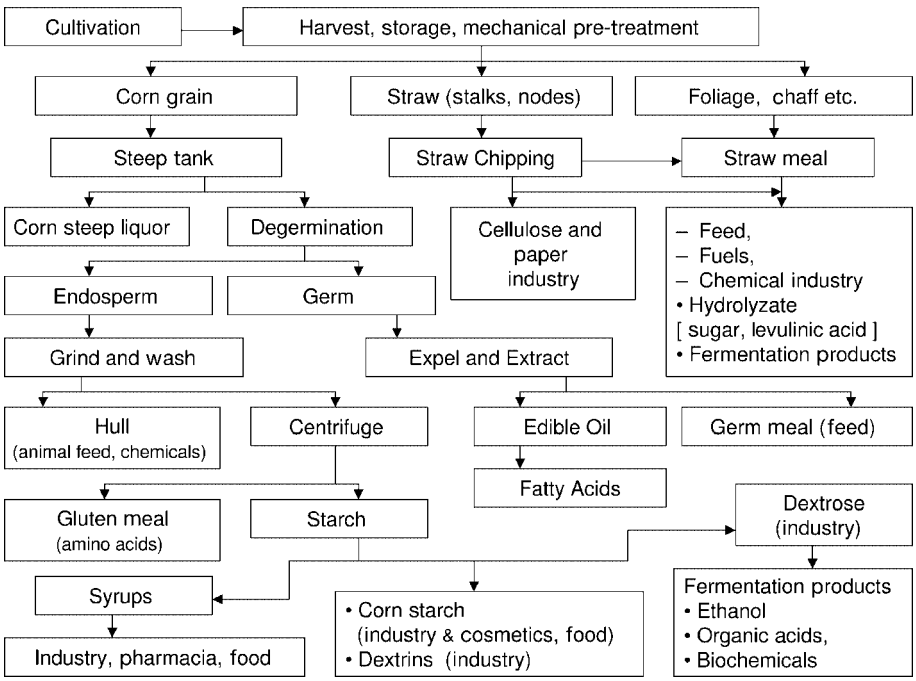

Figure 15. Products from a whole-crop wet mill-based biorefinery

The advantages of the green biorefinery are high biomass profit per hectare, good coupling with agricultural production, and low price of the raw materials. Simple technology can be used and there is good biotechnical and chemical potential for further conversion (Fig. 16). Rapid primary processing or use of preservation methods, for example silage production or drying, are necessary, both for the raw materials and the primary products, although each method of preservation changes the content of the materials.

In addition to cellulose and starch, the press cake contains valuable dyes and pigments, crude drugs and other organic compounds. The green juice contains proteins, free amino acids, organic acids, dyes, enzymes, hormones, other organic substances, and minerals. Application of the methods of biotechnology results in conversion, because the plant water can simultaneously be used for further treatment. In addition, the lignin–cellulose composite is not as intractable as lignocellulose-feedstock materials. Starting from green juice the main focus is directed toward products such as lactic acid and its derivatives, amino acids, ethanol, and proteins. The press cake can be used for production of green feed pellets, as raw material for production of chemicals, for example levulinic acid, and for conversion to syngas and hydrocarbons (synthetic biofuels). The residues of substantial conversion are suitable for production of biogas combined with generation of heat and electricity (Fig. 17). Reviews have been published on the concepts, contents, and goals of the green biorefinery [76, 78, 124, 146].

5.5. Two-platform Concept and Syngas

The "two-platform concept" is one which uses biomass consisting, on average, of 75 % carbohydrates which can be standardized as a "intermediate sugar platform", as a basis for further conversion, but which can also be converted thermochemically into synthesis gas and the products made from this. The "sugar platform" is based on biochemical conversion processes and focuses on fermentation of sugars extracted from biomass feedstocks. The "syngas platform" is based on thermochemical conversion processes and focuses on the gasification of biomass feedstocks and by-products from conversion processes [66, 101, 139]. In addition to the gasification other thermal and thermochemi-

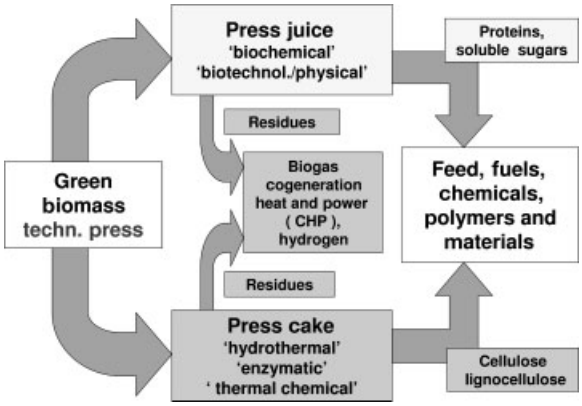

Figure 16. A "green biorefinery" system

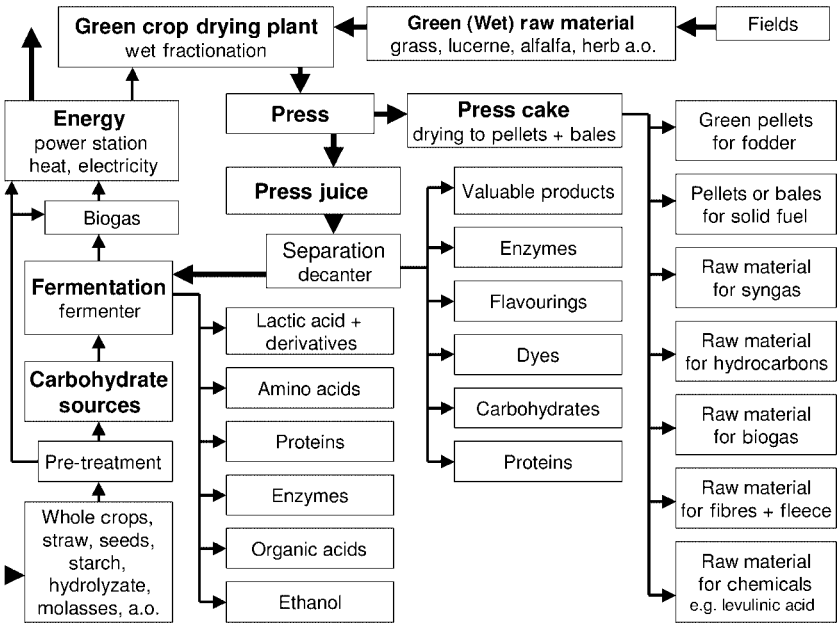

Figure 17. Products from the green biorefinery. In this illustration a green biorefinery has been combined with a green crop-drying plant [81, 82]

cal biomass conversion methods have also been described — hydrothermolysis, pyrolysis, thermolysis, and burning. The application chosen depends on the water content of biomass [174].

The gasification and other thermochemical conversions concentrate on utilization of the precursor carbohydrates and their intrinsic carbon and hydrogen content. The proteins, lignin, oils and lipids, amino acids, and other nitrogen and sulfur-containing compounds occurring in all biomass are not taken into account (Fig. 18).

The advantage of this concept is that the production of energy, fuels, and biobased products is possible using only slightly complex and low-tech technology, for example saccharification and syngas technology [3]. The sugar platform also enables access to a huge range of family-tree-capable chemicals (Figs. 7 and 8).

In-situ conversion of biomass feedstock into liquid or gas could be one way of using existing infrastructure (developed pipe network), but with the disadvantages of the need to remove

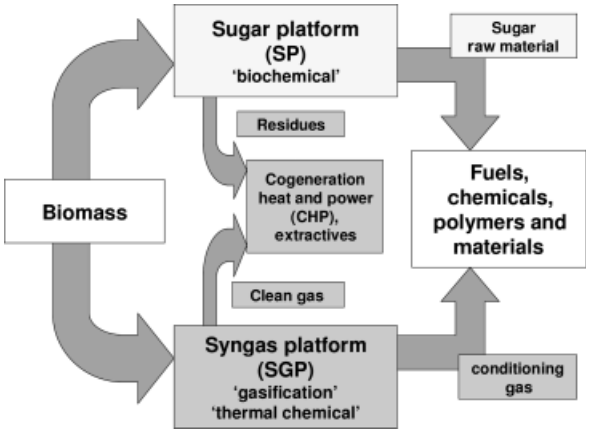

Figure 18. The sugar platform and the syngas platform [139]

hetero-atoms (O, N, S) and minerals present in the biomass and the highly endothermic nature of the syngas process [175]. Currently, production of simple alcohols, aldehydes, mixed alcohols, and Fischer-Tropsch liquids from biomass is not economically viable and additional developments are required [101] (Fig. 19).

Recent research approaches for the production of liquid fuels are aqueous-phase catalytic processes for the conversion of biobased building blocks into H and C1–C15 alkanes [176].

6. Biorefinery Economy

Industrial utilization of raw materials from agriculture, forestry, landscape conservation, and organic waste management for large-scale converting industry and chemical industry as well for producing energy is still in its beginnings. The concept of a biorefinery is analogous to a petroleum refinery where a feedstock, crude oil, is converted to fuels, chemical products, and energy. In the case of a biorefinery, plant biomass is used as a feedstock to produce a diverse set of products for all areas of life, such as fuels, chemicals, polymers, lubricants, adhesives, fertilizer, and energy [132, 133, 177].

Although similar to oil refineries, biorefineries exhibit some important differences. First, biorefineries can utilize a variety of feedstocks. Consequently, they require the development of a larger range of processing technologies and biorefinery systems to deal with compositional differences in the feedstock. Second, the biomass feedstock is bulkier (has a lower energy density) than fossil fuels. Therefore, economics dictate decentralized biorefineries closer to feedstock sources. Conversion of lignocellulosic feedstock to the main products ethanol and furfural has been conceived with optimal profitability (see 5.2, [77]).

The economics of biorefineries are dependent on the production of co-products, such as protein, chemicals, polymers, and power, to provide revenue streams that offset processing costs. Generation of co-products also results in more efficient utilization of biomass and land along with more effective use of invested capital [178, 179]. Generally, because of low cost, plentiful supply, and amenability to biotechnology, carbohydrates appear likely to be the dominant source of feedstock for biocommodity processing [180].

Biorefineries have the potential to reduce sugar costs from about 8–9 cents per pound in the USA today (ca. 2006) to about 4 cents per pound in three to four years, and even lower as the biorefineries become more integrated. It is even possible to foresee a case in which the net cost of producing sugar in an integrated biorefinery is zero, because byproducts such as lignin and proteins generate the value [181].

Plants like NatureWorks PLA are Generation-II biorefineries (see 5.5 [1, 28, 182]). Polylactic acid (PLA) is current produced by using corn as feedstock. The corn-based sugar raw material can account for approximately 25–40 % of the total cost of PLA. Today PLA made from corn is competitive with traditional

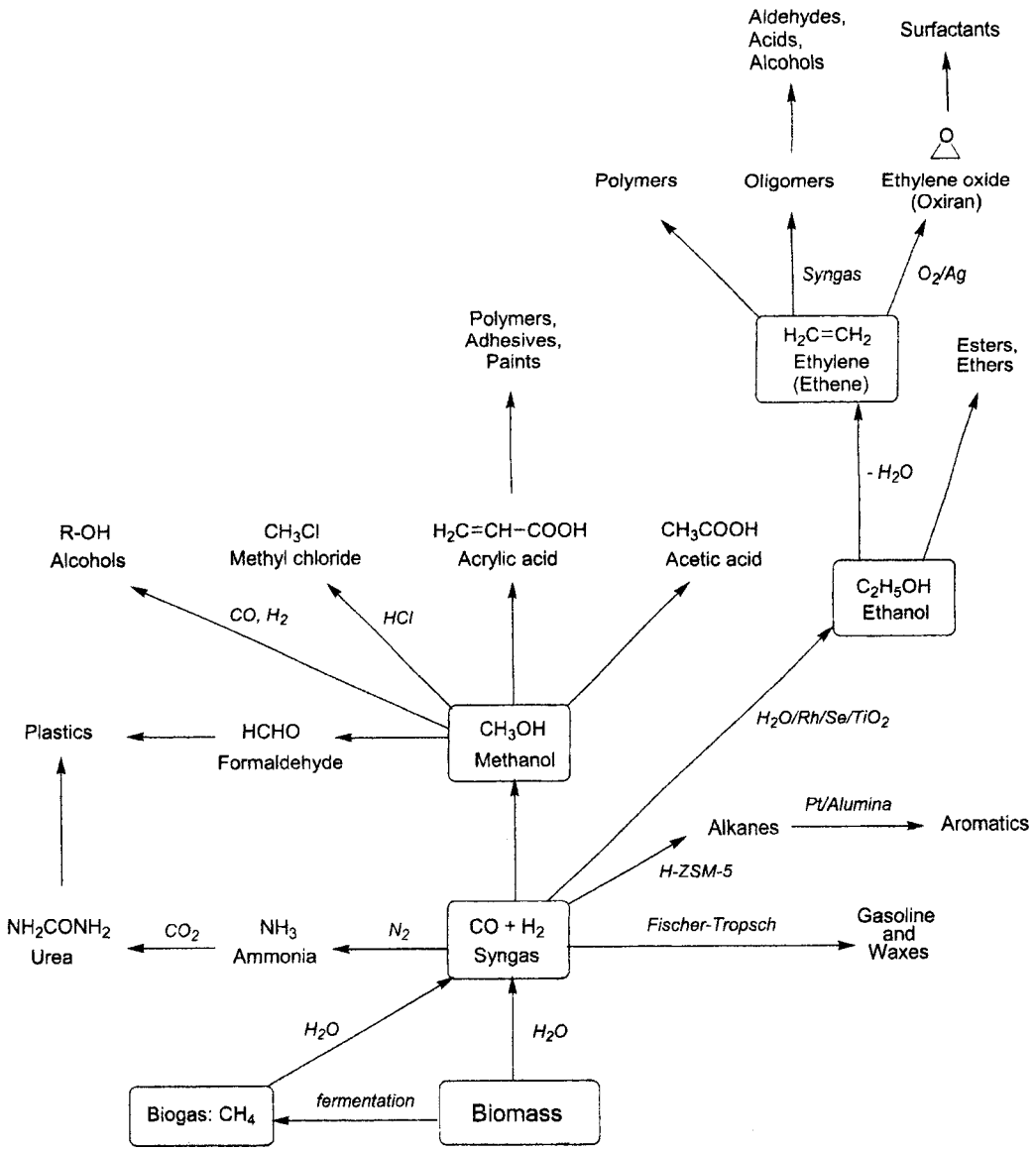

Figure 19. Syngas-based product family tree [175, 190]

petrochemical-based polymers, given its combination of performance in applications, costs, and environmental benefits. Lignocellulose biorefineries are important to PLA because of the improvements in both the cost of the raw material and the overall environmental footprint [1].

In a lignocellulosic biorefinery in which multiple value-added co-products are produced and the lignin is used for energy, the cost of PLA would be lower, primarily because of the lower cost of sugars. In this case the cost of sugar raw materials could be as low as 15–25 % of the total cost of PLA production. Importantly, the environmental footprint would improve because of the replacement of petrochemical-based energy by lignin-based renewable energy to drive the manufacturing process. PLA made by a lignocellulosic biorefinery in the future will have competitive advantages.

Biorefineries utilizing lignocellulosic feedstocks hold the promise of lower net feedstock

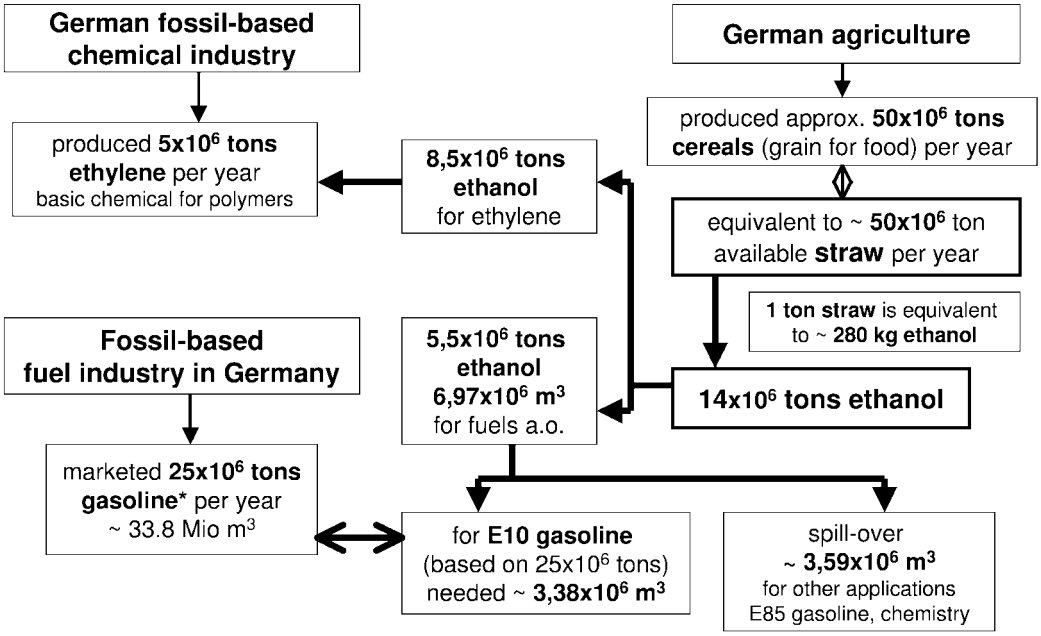

Figure 20. Exemplary calculation on the use of straw for the production of ethanol for fuels and basic chemicals in Germany

costs combined with an improved environmental footprint. PLA already has an advantage over some traditional petrochemical-based polymers such as poly(ethylene terephthalate), PET. Improvements in cost will make it competitive with polypropylene and polystyrene, too.

Recently (ca. 2006), the goal of the U.S. Department of Agriculture and the U.S. Department of Energy is the additional supply of 10^9 tons of biomass at a price of 35 $ per ton (dry matter) per year for industrial chemical and biotechnological utilization, without the restriction of today's applications of biomass from agriculture and forestry [183]. Annually in the USA, 400×10^6 tons of straw (dry matter) has been produced but not used [146, 184].

Example of Use of Straw for Chemicals and Fuels in Germany. Germany produces on average 50×10^6 t/a (dry matter) of cereal grains. The ratio of grain to straw (chaff, nodes, ears, leaves) is approx. 1:1.3 [155].

That means that approximately 50×10^6 t (dry matter) of straw can be used for industry without reducing today's use in organic fertilizing and natural field cultivation. Furthermore, a life-cycle analysis has shown that the use of straw for a lignocellulosic biorefinery is more ecologically advantageous than other possible applications (field cultivation, production of liquid fuels via Fischer–Tropsch, or energy production) [185].

An exemplary calculation for using of this straw for industry (basic chemicals) and fuels (ethanol production) is shown in Figure 20.

The generated ethanol could cover both the demand for ethylene (fossil-based production in Germany: 5×10^6 t) and for E10 fuels (German consumption of gasoline/Otto petrol: 25×10^6 t).The residual ethanol could be applied as a chemical or for E85 fuel (Fig. 20).

7. Outlook and Perspectives

Biorefineries are production plants in which biomass is economically and ecologically converted to chemicals, materials, fuels, and energy. For successful development of "industrial biorefinery technologies" and "biobased products" several problems must be solved. It will be necessary to increase the production of substances (cellulose [186], starch, sugar [187], oil) from basic biogenic raw materials and to promote

the introduction and establishment of biorefinery demonstration plants [188, 189]. Ecological transport of biomass must also be developed, for example, utilization of already developed pipelines.

Another important aspect is committing chemists, biotechnologists, and engineers to the concept of biobased products and biorefinery systems and promoting the combined biotechnological and chemical conversion of substances. Last, but not least, the development of systematic approaches to new syntheses and technologies is required to meet the sustainable principles of "ideal synthesis" and "principles of green chemistry and process engineering" [191 – 193].

The biobased economy and its associated biorefineries will be shaped by many of the same forces that shaped the development of the hydrocarbon economy and its refineries over the past century [182, 194].

Many of the biorefinery elements required for financial success seem to be currently present, for example, high-volume, low-cost applications (ethanol, lactic acid), multiple alternative product streams, ability to shift to different products quickly when required [195], and acceptable cost structure for biomass.

Several financial requirements will enable the biorefinery concept to become a reality in the near future:

- Parts of the overall biorefinery scheme that can operate in a stand-alone manner from a financial perspective, and the ability to evolve from there
- The potential for multiproduct or co-product streams, which may ultimately drive efforts to engineer crops to maximize their potential
- The ability to make use of a variety of feedstocks as hedging strategy and the advantage of pricing opportunities:
- Manageable capital requirements and favorable investment environment
- Government enablement (funding demonstration/development pilot plants, reasonable regulation of biorefinery operations, favorable taxation, strategic market support to give new biorefinery products a head start [196].

8. References

1. Gruber, P.; Henton, D.E.; Starr, J.; Polylactic Acid from Renewable Resources, [B. Kamm, P.R. Gruber, M. Kamm (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*] Vol. 2, 381–407, Wiley-VCH, 2006
2. Ragauskas A. J; Williams Ch. K; Davison B. H; Britovsek G.; Cairney J.; Eckert Ch. A; Frederick W. J Jr; Hallett J. P; Leak D. J; Liotta Ch. L; Mielenz J. R; Murphy R.; Templer R.; Tschaplinski T.; The path forward for biofuels and biomaterials. *Science* **311**(5760), (2006) 484–489.
3. Jong, E. de; Ree, R. van; Tuil, R. van; Elbersen, W.; Biorefineries for the Chemical Industry, [B. Kamm, P.R. Gruber, M. Kamm (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*] Vol. 1, 85–111, Wiley-VCH, 2006
4. National Research Council: *Biobased Industrial Products, Priorities for Research and Commercialization*, National Academic Press, Washington D.C. 2000, a) 74, ISBN 0-309-05392-7.
5. A. S. Marggraf: *Histoire de l'Acadmie Royale des Sciences et Belles Lettres, Ann 1748*, Preußische Akademie, Berlin 1749.
6. A. S. Marggraf: *Chym. Schriften (Chemische Schriften)*, Bd. 2, Berlin 1761–1767.
7. G. S. C. Kirchhoff, *Schweigers Journal für Chemie und Physik*, **4** (1812) 108.
8. C. Graebe: *History of Organic Chemistry*, Bd 1, Springer, Berlin 1920, a) 28, b) 122f.
9. N. Kosaric, F. Vardar-Sukan: "Potential Sources of Energy and Chemical Products" in M. Roehr (ed.): *The Biotechnology of Ethanol*, Wiley–VCH, Weinheim 2001, p. 132, ISBN 3-527-30199-2.
10. S. C. Prescott, C. G. Dunn: *Industrial Microbiology*, 3rd ed., McGraw-Hill, New York 1959.
11. D. Osteroth: *From Coal to Biomass*, Springer, New York 1989, p. 192ff, ISBN 0-387-50712-4.
12. W. J. McKillip: "Furan and Derivatives" in: *Ullmann's Encyclopedia of Industrial Chemistry*, 6th ed., vol. 15, Wiley–VCH, Weinheim 2003, p. 187ff.
13. www.furan.com
14. H. Pringsheim, *Cellulosechemie* **2** (1921) 123.
15. F. Ullmann: *Enzyklopädie der Technischen Chemie*, 2. Aufl., 5. Bd., Urban und Schwarzenberg, Berlin 1930, p. 442ff.

16. A. Payen: *Comptes rendus de l'Acadmie des Sciences (C. r.)* **8** (1839) 51.
17. E. Gruber, T. Krause, J. Schurz: "Cellulose" in: *Ullmann, Enzyklopädie der technischen Chemie*, 4. Aufl. Bd. 9, Verlag Chemie, Weinheim 1975, p. 184–191.
18. G. J. Mulder, *J. Prakt. Chem.* **21** (1840) 219.
19. A. E. Staley, Mfg. Co. A.E. (Decatur Ill.); Levulinic Acid 1942 [C.A. 36, 1612].
20. Kitano et al.: "Levulinic Acid, A New Chemical Raw Material – Its Chemistry and Use", *Chemical Economy and Engineering Review* (1975), 25–29.
21. R. H. Leonard: "Levulinic acid as a Basic Chemical Raw Material", *Ind. Eng. Chem.* (1956), 1331–1341.
22. W. Norman: Process for Converting Unsaturated Fatty Acids or their Glycerides into Saturated Compounds, BP 1515, 1903.
23. W. Sandermann: *Grundlagen der Chemie und chemischen Technologie des Holzes*, Akademische Verlagsgesellschaft, Leipzig 1956, p. 147.
24. F. Tiemann, W. Haarmann: "Ueber das Coniferin und seine Umwandlung in das aromatische Princip der Vanille", *Ber. dt. chem. Ges.* **7** (1874) 608–623.
25. M. J. W. Dignum, J. Kerler, R. Verpoorte: "Vanilla Production: Technological, Chemical and Biosynthetic Aspects", *Food Rev. Intern.* **17** (2001) 199–219.
26. L. Vaupel: "Vanille and Vanillin", *Pharm. Zeit.* **38** (2002).
27. H. Carothers, G. L. Dorough, F. J. Van Natta, *J. Am. Chem. Soc.* **54** (1932) 761.
28. P. R. Gruber, M. O'Brien: "Polylactides "Nature Works" PLA" in: Y. Doi, A. Steinbüchel (eds.): *Biopolymers, Polyester III*, Wiley-VCH, Weinheim 2002.
29. P. Walden in: C. Graebe (ed.): *History of Organic Chemistry since 1880*, Bd. 2, Springer, Berlin 1941, p. 686.
30. W. R. Pötsch: *Lexicon of famous Chemists*, Bibliographisches Institut, Leipzig 1988, a) 305, ISBN 3-323-00185-0.
31. Borregaard LignoTech; marathon co.; http://www.ltus.com.
32. B. W. Peckham: "The First Hundred Years of Corn Refining in the United States", *Corn Annual 2000*, Corn Refiners Association, Washington 2000.
33. D. L. Johnson: "The Corn Wet Milling and Corn Dry Milling Industry — A Base for Biorefinery Technology Developments" in B. Kamm, P. R. Gruber, M. Kamm (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*, vol. 1, Wiley-VCH, Weinheim 2006, pp. 345–353.
34. ICI: Improvement in or related to the utilization of grass and other green crops, Brit. Pat. BP 511,525, 1939 (R. E. Slade, J. H. Birkinshaw).
35. N. W. Pirie, *Chem. Ind.* **61** (1942) 45.
36. N. W. Pirie, *Nature* **149** (1942) 251.
37. W. Heier, *Grundlagen der Landtechnik* **33** (1983) 45–55.
38. B. Kamm, M. Kamm: "The Green Biorefinery – Principles, Technologies and Products", *Proceed. 2nd Intern. Symp. Green Biorefinery*, October 13–14, 1999, Feldbach, Austria, 1999, S. 46–69.
39. B. E. Knuckles, E. M. Bickoff, G. O. Kohler, *J. Agric. Food Chem.* **20** (1972) 1055.
40. F. M. Schertz, *Ind. Eng. Chem.* **30** (1938) 1073–1075.
41. W. H. Shearon, O. F. Gee, *Ind. Eng. Chem.* **41** (1949) 218–226.
42. M. A. Judah, E. M. Burdack, R. G. Caroll, *Ind. Eng. Chem.* **46** (1954) 2262–2271.
43. W. J. Hale: *The Farm Chemurgic*, The Stratford Co., Boston 1934.
44. C. Borth: *Pioneers of Plenthy*, Bobbs-Merril Co, Indianapolis, New York 1939.
45. D. L. Lewis: *The Public Image of Henry Ford*, Wayne State University Press, Detroit 1976.
46. E. N. Brandt: *Growth Company – Dow Chemical's First Century*, Michigan State University press, East Lansing 1997.
47. F. Bergius; *Trans. Inst. Chem. Eng. (London)* **11** (1933) 162.
48. H. Scholler: PhD thesis, Technical University of Munich, 1923.
49. H. Scholler: F.P. 777,824, 1935.
50. R. Katzen, G. T. Tsao: "A View of the History of Biochemical Engineering", *Advances in Biochemical Engineering/Biotechnology*, vol. 70, Springer, Berlin 2000.
51. T. F. Conrad: "Holzzuckerbrennerei" in: R. Reiff et al. (eds.): *Die Hefen*, vol. II 8F., Verlag Hans Carl, Nürnberg 1962, p. 437–444.
52. S. C. Prescott, C. D. Dunn: *Industrial Microbiology*, 3rd ed., McGraw-Hill, New York 1959.
53. E. E. Harris et al.: "Madison Wood Sugar Process", *Ind. Eng. Chem.* **38** (1946) 896–904.
54. T. E. Timell, *Tappi*, **44** (1961) 99.
55. A. J. Stamm: *Wood and Cellulose Science*, Ronald Press, New York 1964.

56. R. L. James: *The pulping of wood*, 2nd ed., vol. 1, McGraw-Hill, New York 1969, p. 34.
57. D. L. Brink, A. A. Pohlmann, *Tappi* **55** (1972) 381ff.
58. M. Oshima: *Wood Chemistry – Process Engineering Aspects*, Noyes development Corp., New York 1965.
59. J. Puls, H. H. Dietrichs: "Energy from Biomass", *Proceed. of First European Comm. Inter. Conf. on Biomass*, Brighton, UK, November 1980, p. 348, ISBN: 0-85334-970-3.
60. S. Y. Shen: "Wood Grass Production Systems for Biomass", *Proceed. Midwest Forest Economist Meeting*, Duluth, Minnesota, April 1982.
61. S. Y. Shen: "Biological Engineering for Sustainable Biomass Production" in: E.O. Wilson (ed.): *Biodiversity*, National Academy of Sciences/Smithsonian Institution, National Academic Press, Washington DC 1988; German Edition: *Ende der biologischen Vielfalt?*, 1992, p. 404–416, ISBN 3-89330-661-7.
62. R. Carlsson: "Trends for future applications of green crops", *Forage Protein Conservation and Utilization*, Proceed. of EFC Conf. 1982, Dublin, Ireland, 1982, p. 57–81.
63. R. Carlsson: "Green Biomass of Native Plants and new Cultivated Crops for Multiple Use: Food, Fodder, Fuel, Fibre for Industry, Photochemical Products and Medicine" in: G.E. Wickens, N. Haq, P. Day (eds.): *New Crops for Food and Industry*, Chapman and Hall, London 1989.
64. B. E. Dale: "Biomass refining: protein and ethanol from alfalfa", *Ind. Eng. Product Research and Development* **22** (1983) 446.
65. A. J. Hacking: "The American wet milling industry", *Economic Aspects of Biotechnology*, Cambridge University Press, New York 1986, p. 214–221.
66. D. H. White, D. Wolf: *Research in Thermochemical Biomass*, Elsevier Applied Science, New York 1988.
67. E. S. Lipinsky: "Chemicals from Biomass: petrochemical substitution options", *Science* **212** (1981) 1465–1471.
68. D. L. Wise (ed.): *Organic Chemical from Biomass*, The Benjamin/Cummings Publishing Co., Inc., Menlo Park, California, 1983.
69. J. E. Bailey, D. F. Ollis: *Biochemical Engineering Fundamentals*, 2nd ed., McGraw-Hill, New York 1986.
70. H. H. Szmant: *Industrial Utilization of Renewable Resources*, Technomic Publishing, Lancaster, Pa., 1987.
71. M. Eggersdorfer, J. Meyer, P. Eckes: "Use of renewable resources for non-food materials", *FEMS Microbiolol. Rev.* **103** (1992) 355–364.
72. D. J. Morris, I. Ahmed: *The carbohydrate Economy: Making Chemicals and Industrial Materials from Plant Matter*, Institute of Local Self Reliance, Washington D.C. 1992.
73. J. J. Bozell, R. Landucci: *Alternative feedstock program – technical and economic assessment*, US Department of Energy, 1992.
74. L. B. Schilling: "Chemicals from alternative feedstock in the United States", *FEMS Microbiol. Rev.* **16** (1995) 1001–1110.
75. L. B. de la Ross et al.: "An integrated process for protein and ethanol from coastal Bermuda grass", *Appl. Biochem. Biotechnol.* **45/46** (1994) 483–497.
76. K. Soyez, B. Kamm, M. Kamm (eds.): "The Green Biorefinery", *Proceedings of 1st International Green Biorefinery Conference, Neuruppin, Germany, 1997*, GÖT, Berlin 1998, ISBN 3-929672-06-5, German and English.
77. D. L. Van Dyne, M. G. Blas, L. D. Clements: "A strategy for returning agriculture and rural America to long-term full employment using biomass refineries" in: J. Janeck (ed.): *Perspectives on new crops and new uses*, ASHS Press, Alexandria, Va., 1999, p. 114–123.
78. M. Narodoslawsky: "The Green Biorefinery", Proceedings 2nd Intern. Symp. Green Biorefinery, Feldbach, Austria, 1999.
79. R. V. Nonato, P. E. Mantellato, C. E. V. Rossel: "Integrated production of biodegradable plastic, sugar and ethanol" *Appl. Microbiol. Biotechnol.* **57** (2001) 1–5.
80. H. Ohara: "Biorefinery", *Appl. Microb. Biotechn. (AMB)* **62** (2003) 474–477.
81. B. Kamm, M. Kamm: "Principles of Biorefineries", *Appl. Microbiol., Biotechnol., (AMB)* **64** (2004) 137–145.
82. B. Kamm, M. Kamm: "Biorefinery-Systems", *Chem. Biochem. Eng. Q.* **18** (2004) 1–6.
83. U.S. Department of Energy (DEO): National Biomass Initiative and Energy, Environmental and Economics (E3) Handbook, www.bioproducts-bioenergy.gov.
84. L. R. Lynd, J. H. Cushman, R. J. Nichols, C. E. Wyman: "Fuel Ethanol from Cellulosic Biomass", *Science* **251** (1991) 1318.

85. F. A. Keller: "Integrated bioprocess development for bioethanol production" in: C.E. Wyman (ed.): *Handbook ob bioethanol: production and utilization*, Taylor and Francis, Bristol 1996, p. 351–379.
86. C. E. Wyman: *Handbook on Bioethanol: Production and Utilization*, Applied Energy Technology Series, Taylor and Francis, Bristol 1996.
87. L. Lynd: "Overview and evaluation of fuel ethanol from cellulosics biomass: technology, economics, the environment, and policy", *Annual Review of Energy and the Environment* **21** (1996) 403–465.
88. M. Galbe, G. Zacci: "A review of the production of ethanol from softwood", *Appl. Microbiol. Biotechnol.* **59** (2002) 618–628.
89. R. Datta et al.: "Technological and economics potential of poly (lactic acid) and lactic acid derivatives" *FEMS Microbiol. Rev.* **16** (1995) 221–231.
90. U. Witt, R. J. Müller, H. Widdecke, W.-D. Deckwer: "Synthesis, properties and biodegradability of polyesters based on 1,3-propanediol", *Macrom. Chem. Phys.* **195** (1994) 793–802.
91. T. Yosida, T. Nagasawa: "ε-Poly-L-lysine: microbial production, biodegradation and application potential", *Appl. Microbiol. Biotechnol.* **62** (2003) 21–26.
92. C. Potera: "Genencor and DuPont create"green" polyester", *Genet. Eng. News* **17** (1997) 17.
93. Y. Poirier, C. Nawrath, C. Somerville: "Production of polyhydroxyalkanoates, a family of biodegradable plastics", *Bio/technology* **13** (1995) 142–150.
94. S. Warwel et al.: "Polymers and Surfactants on the basis of renewable resources", *Chemosphere* **43** (2001) 39–48.
95. J. J. Bozell: "Alternative Feedstocks for Bioprocessing" in: R.M. Goodman (ed.): *Encyclopedia of Plant and Crop Science*, Dekker, New York 2004, ISBN:0-8247-4268-0.
96. C. Webb, A. A. Koutinas, R. Wang: "Developing a Sustainable Bioprocessing Strategy Based on a Generic Feedstock", *Adv. Biochem Eng./Biotechn.* **87** (2004) 195–268.
97. J. Gravitis, M. Suzuki: "Biomass Refinery – A Way to produce Value Added Products and Base for Agricultural Zero Emissions Systems" in: *Proc. 99 Intern. Conference on Agric. Engineering, Beijing, China 1999*, United Nations University Press, Tokyo 1999, III-9–III-23.
98. D. L. Van Dyne et al.: *Estimating the Economic Feasibility of Converting Ligno-Cellulosic Feedstocks to Ethanol and Higher Value Chemicals under the refinery concept: A Phase II Study*, University of Missouri, 1999, OR22072-58.
99. Z. Kurtanjek, *Chemical and Biochemical Engineering Quarterly, Special Issue* **18** (2004) 1–88.
100. J. J. Marano, J. L. Jechura: "Biorefinery Optimization Tools – Development and Validation" in: 25th. Symposium on Biotechnology for Fuels and Chemicals: Program and Abstracts, National Renewable Energy Laboratory, Golden, CO, No. NREL/BK-510-33708, 2003, p. 104.
101. T. Werpy, G. Petersen (eds.): Top Value Added Chemicals from biomass, U.S. Department of Energy, Office of scientific and technical information, 2004, No.: DOE/GO-102004-1992, www.osti.gov/bridge.
102. O. J. M. Goddijn: "Plants as bioreactors", *Trends Biotechnol.* **13** (1995) 379–387.
103. D. Wilke: "Chemicals from biotechnology: molecular plant genetic will challenge the chemical and the fermentation industry", *Appl. Microbiol. Biotechnol., (AMB)* **52** (1999) 135–145.
104. R. Anex, *Journal of Ind. Ecology, Special Issue* **7** (2003) 1–235.
105. R. G. Ludgar; R. J. Woolsey; The new petroleum. *Foreign Affairs* **78** (1999) 88–102.
106. C. E. Wyman: "Production of Low Cost Sugars from Biomass: Progress, Opportunities, and Challenges" in: R.P. Overend and E. Chornet (eds.): *Biomass: A Growth Opportunity in Green Energy and Value-Added Products*, Pergamon Press, Oxford 1999, p. 867–872.
107. R. Bachmann, E. Bastianelli, J. Riese, W. Schlenzka: "Using plants as plants. Biotechnology will transform the production of chemicals", *The McKinsey Quarterly* **2** (2000) 92–99.
108. J. R. Hettenhaus, B. Wooley: Biomass Commercialization: Prospect in the Next 2 to 5 Years, NREL, Golden Colorado 2000, No. NREL/ACO-9-29-039-01.
109. J. Woolsey: "Hydrocarbons to Carbohydrates, The strategic Dimension", *The Biobased Economy of the 21st Century: Agriculture Expanding into Health, Energy, Chemicals, and Materials*, NABC Report 12, National

Agricultural Biotechnology Council, Ithaca, New York 2000, No. 14853.
110. A. Eaglesham, W. F. Brown, R. W. F. Hardy (eds.): *The Biobased Economy of the 21st Century: Agriculture Expanding into Health, Energy, Chemicals, and Materials, NABC Report 12*, National Agricultural Biotechnology Council, Ithaca, New York 2000, No. 14853.
111. US President: Developing and Promoting Biobased Products and Bioenergy, Executive Order 13101/13134, William J. Clinton, The White House, Washington D.C. 1999.
112. US Congress: Biomass Research and Development, Act of 2000, Washington D.C. 2000.
113. Biomass R& D, Technical Advisory Committee; Vision for Bioenergy and Biobased Products in the United States, Washington D.C. Oct. 2002; www.bioproducts-bioenergy.gov/pdfs/BioVision_03_Web.pdf.
114. Biomass R& D, Technical Advisory Committee; Roadmap for Biomass Technologies in the United States, Washington D.C., Dec. 2002; www.bioproducts-bioenergy.gov/pdfs/FinalBiomassRoadmap.pdf.
115. Biotechnology Industrial Organisation: World Congress on Industrial biotechnology and Bioprocessing; http://www.bio.org/worldcongress.
116. Biomass Conferences of the Americas; http://www.nrel.gov/bioam/
117. R. P. Overend, E. Chornet (ed.): "Biomass, a growth opportunity in green energy and value-added products", Proceedings of the 4th Biomass Conference of the Americas, Oakland, CA, Aug. 29 – Sept. 2, 1999, Elsevier, ISBN: 0080430198.
118. Green and Sustainable Chemistry Congress; http://www.chemistry.org.
119. J. J. Bozell: "Feedstocks for the future: using technology development as a guide to product identification", *ACS Symp. Series* **921** (2006) 1–12.
120. Biorefinica – International Symposia Biobased Products and Biorefineries; www.biorefinica.de.
121. B. Kamm, M. Hempel, M. Kamm (eds.): "biorefinica 2004, International Symposium Biobased Products and Biorefineries", *Proceedings and Papers*, October, 27 and 28, 2004, biopos, Teltow 2004, ISBN 3-00-015166-4.
122. H. Zoebelin (ed.): *Dictionary of Renewable Resources*, Wiley-VCH, Weinheim 2001.
123. L. E. Manzer: "Biomass derivatives: a sustainable source of chemicals", *ACS Symp. Series* **921** (2006) 40–51.
124. B. Kamm et al.: "Green Biorefinery Brandenburg, Article to development of products and of technologies and assessment", *Brandenburgische Umweltberichte* **8** (2000) 260–269.
125. European parliament and Council: Directive 2003/30/EC on the promotion of the use of biofuels or other renewable fuels for transport, Official Journal of the European Union L123/42, 17. 05. 2003, Brussels 2003.
126. Gesetz für den Vorrang erneuerbarer Energien: Erneuerbare Energiegesetz, EEG/EnWGuaÄndG., 29. 03.2000, BGBI, 305, 2000.
127. European parliament and Council; Green Paper "Towards a European strategy for the security of energy supply" KOM2002/321, 26. 06. 2002.
128. Umweltbundesamt: *Klimaschutz durch Nutzung erneuerbarer Energien*, Report 2, Erich Schmidt Verlag, Berlin 2000.
129. BioVision2030-Group: Strategiepapier Industrielle stoffliche Nutzung von Nachwachsenden Rohstoffen in Deutschland, Nov. 2003; www.biorefinica.de/bibliothek.
130. Deutscher Bundestag: Rahmenbedingungen für die industrielle stoffliche Nutzung von Nachwachsenden Rohstoffen in Deutschland schaffen, Antrag 15/4943, Berlin 2005.
131. R. Busch et al.: "Biomasse-Industrie – Wie aus "Bio" Chemie wird", *Nachrichten aus der Chemie* **53** (2005) 130–134.
132. European Technology Platform for Sustainable Chemistry, Industrial Biotechnology Section, 2005; www.suschem.org.
133. US Department of Energy (DOE), 1st International International Biorefinery Workshop, July 20/21st 2005, Washington D.C.; www.biorefineryworkshop.com.
134. H. Röper: "Perspektiven der industriellen Nutzung nachwachsender Rohstoffe, insbesondere von Stärke und Zucker", *Mitteilung der Fachgruppe Umweltchemie und Ökotoxikologie der Gesellschaft Deutscher Chemiker* **7** (2001) no 2, 6–12
135. F. W. Lichtenthaler: "The Key Sugars of Biomass: Availability, Present Non-Food Uses and potential Future Development Lines" in B. Kamm et al. (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and*

Future Directions, vol. 2, Wiley-VCH, Weinheim 2006, p. 3–59.
136. B. Kamm, M. Kamm, K. Richter: "Entwicklung eines Verfahrens zur Konversion von hexosenhaltigen Rohstoffen zu biogenen Wirk- und Werkstoffen – Polylactid aus fermentiertem Roggenschrot über organische Aluminiumlactate als alternative Kuppler biotechnischer und chemischer Stoffwandlungen" in: P. B. Czedik-Eysenberg (ed.): *Chemie nachwachsender Rohstoffe*, Österreichisches Bundesministerium für Umwelt (BMUJF), Wien 1997, pp. 83–87.
137. B. Kamm, M. Kamm, M. Schmidt, I. Starke, E. Kleinpeter: "Chemical and biochemical generation of carbohydrates from lignocellulose-feedstock (Lupinus nootkatensis), Quantification of glucose", *Chemosphere* **62** (2006) no 1, 97–105.
138. US Department of Energy; http://www.oit.doe.gov/e3handbook.
139. National Renewable Energy Laboratory (NREL); http://www.nrel.gov/biomass/biorefinery.html.
140. S. Fernando, S. Adhikari, C. Chandrapal, N. Murali: "Biorefineries: Current Status, Challenges, and Future Directions", *Energy Fuels*; 2006; **20**(4) 1727–1737.
141. EuropaBio: *White Biotechnology: Gateway to a more sustainable future*, EuropaBio, Lyon, April 2003.
142. BIO Biotechnology Industry Organisation: New Biotech Tools for a cleaner Environment – Industrial Biotechnology for Pollution Prevention, Re-
source Conservation and Cost Reduction, 2004; http://www.bio.org/ind/pubs/cleaner2004/cleanerReport.pdf.
143. Dti Global Watch Mission Report: Impact of the industrial biotechnology on sustainability of the manufacturing base – the Japanese Perspective, 2004.
144. P. Claus, G. H. Vogel: "Die Rolle der Chemokatalyse bei der Etablierung der Technologieplattform, Nachwachsende Rohstoffe", *Chemie Ingenieur Technik* **78** (2006) no 8, 991–1012 (German).
145. N. Suzuki, H. Yukawa: "Bio-refinery: industrial revolution of the 21st century", *Cellulose Communications* **11** (2004) **4**, 181–187 (Japanese).
146. J. S. Tolan: "Iogen's Demonstration Process for Producing Ethanol from Cellulosic Biomass" in: B. Kamm et al. (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*, vol. 1, Wiley-VCH, Weinheim 2006, pp. 193–208;
147. J. Hettenhaus: "Biomass Commercialization and Agriculture Residue Collection" in: B. Kamm et al. (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*, vol. 1, Wiley-VCH, Weinheim 2006, pp. 317–344.
148. S. Kromus, B. Kamm, M. Kamm, P. Fowler, M. Narodoslawsky: "The Green Biorefinery Concept- Fundamentals and Potential" in: B. Kamm et al. (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*, vol. 1, Wiley-VCH, Weinheim 2006, p. 253–294.
149. B. Kamm, M. Kamm, M. Schmidt, T. Hirth, M. Schulze: "Lignocellulose- based Chemical Products and Product Family Trees" in: B. Kamm et al. (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*, vol. 2, Wiley-VCH, Weinheim 2006, p. 97–149.
150. B. Dale: *Encyclopedia of Physical Science and Technology*, 3rd ed., vol. 2, 2002, p. 141–157.
151. D. J. Hayes, S. Fitzpatrick, M. H. B. Hayes, J. H. R. Ross: "The Biofine Process-Production of levulinic acid from lignocellulosic feedstock" in: B. Kamm et al. (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*, vol. 1, Wiley-VCH, Weinheim 2006, p. 139–164.
152. M. Ringpfeil: Biobased Industrial Products and Biorefinery Systems – Industrielle Zukunft des 21. Jahrhunderts?, 2001; www.biopract.de.
153. J. G. Zeikus, M. K. Jain, P. Elankovan: "Biotechnology of succinic acid production and markets for derived industrial products", *Appl. Microbiol. Biotechnol.* **51** (1999) 545–552.
154. K.-D. Vorlop, Th. Wilke, U. Prüße: "Biocatalytic and catalytic routes for the production of bulk and fine chemicals from renewable resources", in B. Kamm et al. (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*, vol. 1, Wiley-VCH, Weinheim 2006.
155. Die Landwirtschaft 1998 *Lehrbuch für Landwirtschaftsschulen. Band 1. Pflanzliche Erzeugung. 11.* Auflage, München: BLV Verlagsgesellschaft. Münster Hiltrup: Landwirtschaftsverlag. p. 280.
156. G. H. Robertson et al.: "Native or raw starch digestion: a key step in energy efficient

biorefining of grain", *Journal of agricultural and food chemistry* **54** (2006) no 2, 353–365.
157. A. A. Koutinas, R. Wang, G. M. Campbell, C. Webb: "A Whole Crop Biorefinery System: A closed System for the Manufacture of Non-food-Products from Cereal" in B. Kamm et al. (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*, vol. 1, Wiley-VCH, Weinheim 2006, p. 165–191.
158. C. E. V. Rossel, P. E. Mantellato, A. M. Agnelli, J. Nascimento: "Sugar Based Biorefinery — Technology for Integrated Production of Poly(3-hydroxybutyrate), Sugar and Ethanol in: B. Kamm, P. R. Gruber, M. Kamm (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*, vol. 1, Wiley-VCH, 2006, pp. 209–326.
159. A. A. Koutinas, R. Wang, G. M. Campbell, C. Webb: "Developing viable biorefineries for the production of biodegradable microbial plastics and various value-added products", 7th World Congress of Chemical Engineering, Glasgow, United Kingdom, July 10–14, 2005.
160. F. Rexen: New industrial application possibilities for straw. Documentation of Svebio Phytochemistry Group (Danish), Fytokemi i Norden, Stockholm, Sweden, 1986, p. 12.
161. J. Coombs, K. Hall: "The potential of cereals as industrial raw materials: Legal technical, commercial considerations" in: G. M. Campbell, C. Webb, S. L. McKee (eds.): *Cereals – Novel Uses And Processes*, Plenum Publ. Corp., New York 1997, p. 1–12.
162. E. Audsley, J. E. Sells: "Determining the profitability of a whole crop biorefinery" in: G. M. Campbell, C. Webb, S. L. McKee (eds.): *Cereals – Novel Uses and Processes*, Plenum Publ. Corp.; New York 1997, p. 191–294.
163. R. Carlsson: "Sustainable primary production – Green crop fractionation: Effects of species, growth conditions, and physiological development" in: M. Pessarakli (ed.): *Handbook of Plant and Crop Physiology*, Marcel Dekker, New York 1994, pp. 941–963.
164. N. W. Pirie: *Leaf Protein – Its agronomy, preparation, quality, and use*, Blackwell Scientific Publications, Oxford/Cambridge, UK, 1971.
165. N. W. Pirie: *Leaf Protein and Its By-Products in Human and Animal Nutrition*, Cambridge Univ. Press, UK, 1987.
166. R. Carlsson: "Status quo of the utilization of green biomass" in: S. Soyez, B. Kamm, M. Kamm (eds.): *The Green Biorefinery, Proceedings of 1st International Green Biorefinery Conference*, Neuruppin, Germany, 1997, Verlag GÖT, Berlin 1998, ISBN 3-929672-06-5.
167. R. Carlsson: "Food and non-food uses of immature cereals" in: G.M. Campbell, C. Webb, S.L. McKee (eds): *Cereals – Novel Uses and Processes*, Plenum Publ. Corp., New York, USA, 1997, pp. 159–167.
168. R. Carlsson: "Leaf protein concentrate from plant sources in temperate climates" in: L. Telek, H.D. Graham (eds.): *Leaf Protein Concentrates*, AVI Publ. Co., Inc., Westport, Conn., USA, 1983, pp. 52–80.
169. L. Telek, H. D. Graham (eds.): *Leaf Protein Concentrates*, AVI Publ., Co., Inc., Westport, Conn., USA, 1983.
170. R. J. Wilkins: *Green Crop Fractionation*, The British Grassland Society, c/o Grassland Research Institute, Hurley, Maidenhead, SL6 5LR, UK, 1977.
171. I. Tasaki (ed.): "Recent Advances in Leaf Protein Research", Proc. 2nd Int. Leaf Protein Res. Conf., Nagoya, Japan, 1985.
172. P. Fantozzi (ed.): Proc. 3rd Int. Leaf Protein Res. Conf., Pisa-Perugia-Viterbo, Italy, 1989.
173. N. Singh (ed.): *Green Vegetation Fractionation Technology*, Science Publ. Inc., Lebanon, NH 03767, USA, 1996.
174. C. Okkerse, H. van Bekkum: "From fossil to green", *Green Chemistry* **4** (1999) 107–114.
175. M. Lancaster: "The Syngas Economy" in: *Green Chemistry*, The Royal Society of Chemistry, Cambridge, UK, 2002, p. 205, ISBN: 0-85404-620-8.
176. G. W. Huber, J. A. Dumesic: "An overview of aqueous-phase catalytic processes for production of hydrogen and alkanes in a biorefinery", *Catalysis Today* **111** (2006) 1–2, 119–132.
177. B. Kamm et al.: *Biorefineries — Industrial Processes and Products, Status Quo and Future Directions*, vols. 1 and 2, Wiley-VCH, Weinheim 2006.
178. Biotechnology Industrial Organization: Development of Biorefineries, BioCycle, April, 2005.
179. C. E. Wyman: "Economics of a biorefinery for coproduction of succinic acid, ethanol, and electricity", Abstracts of Papers, 221st ACS National Meeting, San Diego, CA, United

States, April 1-5, 2001, BIOT-072, American Chemical Society.
180. L. R. Lynd, C. E. Wyman, T. U. Gerngross: "Biocommodity Engineering", *Biotechnol. Prog.* **15** (1999) 777–793.
181. R. Bachmann, J. Riese: "Industrial Biotech — Setting Conditions to Capitalize on the Economic Potential" in: B. Kamm et al. (eds.): *Biorefineries — Industrial Processes and Products, Status Quo and Future Directions*, vol. 2, Wiley-VCH, Weinheim 2006, pp. 445–462.
182. E. T. H. Vink et al.: "The sustainability of Nature Works polylactide polymers and Ingeo polylactide fibers: An update of the future", *Macromolecular Bioscience* **4** (2004) **60**, 551–564.
183. U.S. Department of Agriculture (USDA) and U.S. Department of Energy (DOE) (eds.): *Biomass as Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-Ton Annual Supply*, U.S. Department of Energy, Office of Scientific and Technical Information, P.O. Box 62, Oak Ridge, TN, 2005.
184. R. Katzen, G. T. Tsao: "A View of the History of Biochemical Engineering", *Advances in Biochemical Engineering/Biotechnology*, vol. 70, Springer, Berlin 2000.
185. Office of Technology Assessement at the German Parliament, Monitoring: Industrial conversion of biomass (in print substantial German Parliament Berlin).
186. W. E. F. Mabee, D. G. Evan, P. N. McFarlane, J. N. Saddler: "Canadian biomass reserves for biorefining", *Applied Biochemistry and Biotechnology* **129-132** (2006) 22–40.
187. L. A. Edye, W. O. S. Doherty, J. A. Blinco, G. E. Bullock: "The sugarcane biorefinery: energy crops and processes for the production of liquid fuels and renewable commodity chemicals", *International Sugar Journal* **108** (2006) no 1285, 19–20, 22–27.
188. T. Eggeman, D. Verser: "The importance of utility systems in today's biorefineries and a vision for tomorrow", *Applied Biochemistry and Biotechnology* **129-132** (2006), 361–381.
189. E. Kendall Pye, "Biorefining; a major opportunity for the sugar cane industry", *International Sugar Journal* **107** 1276, 222–224, 226, 228, 230, 253 (2005).
190. A. S. Matlack: "The Use of Synthesis Gas from Biomass", in: *Introduction to Green Chemistry*, Marcel Dekker, New York 2001, p. 369, ISBN: 0824704118.
191. J. H. Clark: "Green Chemistry. Challenges and opportunities", *Green Chemistry* **1** (1999) 1–8.
192. M. Lancaster: "The Biorefinery" in: *Green Chemistry*, The Royal Society of Chemistry, Cambridge, UK, 2002, p. 207, ISBN: 0-85404-620-8.
193. P. T. Anastas, J. C. Warner: *Green Chemistry. Theory and Practice*, Oxford University Press, New York 1998.
194. B. E. Dale, S. Kim: "Biomass Refining Global Impact – The Biobased Economy of the 21st Century" in: B. Kamm et al. (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*, vol. 1, Wiley-VCH, Weinheim 2006, p. 41–66.
195. V. Kochergin, M. Kearney: "Existing biorefinery operations that benefit from fractal-based process intensification", *Applied Biochemistry and Biotechnology* **129-132** (2006) 349–360.
196. B. Dean, T. Dodge, F. Valle, G. Chotani: "Development of Biorefineries- Technical and Economical Considaration" in: B. Kamm et al. (eds.): *Biorefineries – Industrial Processes and Products, Status Quo and Future Directions*, vol. 1, Wiley-VCH, Weinheim 2006, pp. 67–83.

Cereal Products

CHRISTOPHER RUBIN, Bühler AG, Business Unit Pasta and Extruded Products, Uzwil, Switzerland

1. Introduction 1
2. Food Extrusion Technology 2
3. Pasta 2
3.1. Definition 2
3.2. Market for Pasta 2
3.3. Production 3
3.3.1. Raw Materials 3
3.3.1.1. Semolina 3
3.3.1.2. Water and Additive Addition 3
3.3.2. Material Handling 3
3.3.3. Short Goods, Long Goods, and Process Technology 4
3.3.4. Drying 5
3.3.5. Intermediate Storage 6
3.3.6. Thermal and Electrical Infrastructures 7
3.3.7. Pasta Quality 7
4. Modified Flours 8
4.1. Definition and Possible Applications 8
4.2. Production 8
4.2.1. Process 8
4.2.2. Production Plant for Modified Flours 8
4.3. Technological Aspects of Extrusion Cooking 9
4.4. Effects of the Process Variables on Parameters and Functional Properties of Modified Flours . . 11
4.4.1. Effect of Process Variables on Parameters 11
4.4.2. Effect of Parameters on Functional Properties 11
4.4.3. Effect of Functional Properties on Finished Products 12
5. Breakfast Cereals 12
5.1. Definition and Historical Background 12
5.2. Processes 12
5.2.1. Batch Cooking 12
5.2.2. Gun Puffing 13
5.2.3. Shredding 13
5.2.4. Cooking Extrusion 13
5.2.5. Coating and Inclusions 14
5.2.6. Production Plant for Breakfast Cereals 14
6. Reconstituted and Fortified Cereal Grains, e.g., Rice Kernels . 15
6.1. Importance of Rice as a Food . . . 15
6.2. Rice Milling and Nutrient Losses 16
6.3. Extrusion Process for the Production of Tailor-Made Rice Products 16
6.3.1. Quick-Cooking Rice 17
6.3.2. Instant Rice 17
6.3.3. Additive Product 17
6.3.4. Fortification of Reconstituted Rice . 18
7. References 18

1. Introduction

Food products based on cereal grains belong to the staple foods all over the world. In 2004/2005 2000×10^6 t of cereal grains were harvested [1]. Millions of tons of cereal grains are processed in the plants of the food industry day by day. The most important intermediate product is flour. Grain flour or coarser fractions of the kernels are further processed. One possible process technology is extrusion.

There are many kinds of food products which are produced by extrusion technology such as pasta, breakfast cereals, cereal flakes, baby food, bread crumbs, or binders for, e.g., sauces or soups. In Chapter 2 a short overview on food extrusion technology is given. In the following chapters cereal products are discussed which are produced with the help of extrusion technology. The Chapter 3 describes pasta and the pasta production technology. A description of the extrusion process for the production of modified flour as food ingredients is found in Chapter 4. The Chapter 5 gives an overview over breakfast cereals and their possible production methods. The last chapter 6 is devoted to reconstituted and fortified rice kernels.

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.l06_l01

2. Food Extrusion Technology

Extruders have been used in a large variety and increasing number of food applications. Nearly all of these involve mixing of food ingredients with water to achieve hydration. In most cases, a number of ingredients are involved to make a finished extruded food product. Food ingredients are composed of starches, proteins, fiber, fat, minerals, vitamins and water. Major raw materials are cereal grains, or flours of those, respectively.

Extrusion is generally defined as a process of forcing material through a defined opening. For this to occur, the material must be either completely or partly in a fluid form so that it does not block the opening when operating under reasonable pressure conditions.

In reality, extrusion systems provide other functions including solid conveying, melting, mixing, and metering. In the process the material may undergo chemical as well as physical changes.

In the food industry, single-screw extruders were first used in the pasta industry in the early 1930s. Cooking extruders were introduced in the 1940s to produce expanded corn snacks. Long single-screw cooking extruders were developed in the 1950s for production of pet food and expanded breakfast cereals. Twin-screw corotating cooking extruders were introduced in the late 1970s.

Food extruders have been divided into two general application categories.

For *forming applications*, low-shear single-screw extruders are basically used to form pellets with a minimum energy input to the product. The product applications include pasta, cold formed snacks, and unexpanded, precooked pellets. The screws have a deep channel (height/diameter, i.e., $H/D \approx 4$) with probably no compression and are running at low screw speed (below 50 rpm).

Cooking extrusion applications normally utilize medium- and high-shear single- and twin-screw extruders. In these, significant energy input to the product is provided by viscous heat dissipation with or without heat transfer through the barrel. Their screws run at higher speeds (100–500 rpm) with shallower channel depth and compression ratio up to 5:1 for single-screw extruders. Cooking extruders are used for cooking and forming of expanded or unexpanded products.

3. Pasta

3.1. Definition

The word "pasta" means "dough" in Italian. Its typical ingredients are wheat flour or semolina and water. Alternatives are potato flour (for gnocchi) and maize (corn) or rice flour (gluten-free products). It is also possible to add eggs, natural colorants such as spinach, tomatoes, or vitamins for certain products in the USA. The Italian type of pasta production uses only durum semolina and water. From the point of view of composition, pasta is a very simple food.

3.2. Market for Pasta

Today, a market for "Italian-style" pasta exists worldwide. It is not surprising that of all countries Italy has the higher per capita consumption of pasta. However, as shown in Table 1 in various other countries also high amounts of pasta are consumed.

Table 1. Pasta consumption of different countries [2]

Country	kg pasta per capita and year 2005
Italy	28.0
Venezuela	12.6
Tunisia	11.7
Switzerland	9.8
Greece	9.6
Sweden	9.0
USA	9.0
Peru	8.5
Germany	6.8
UK	2.5
Japan	1.7

As a general rule, pasta consumption has increased strongly since the 1970s. In Italy, on the other hand, consumption has tended to decline somewhat over the past 30 years. Surveys show pasta to be considered as a healthy product that is easy to store and quick to prepare. Besides the US pasta producers, the Italians are also benefiting from the higher pasta consumption in the USA due to the quantities imported from Italy. In the UK, market growth has slowed down to

an annual rate of about 2 % in 2001. The market for fresh pasta is much smaller, but is growing at about 10 % annually.

3.3. Production

The processing of grain into ready-to-consume pasta products involves numerous individual operations. The following chapters are discussing each step from raw material up to storage of dried pasta.

3.3.1. Raw Materials

3.3.1.1. Semolina

Semolina made from durum or hard wheat (for a definition, see → Bread and Other Baked Products, Chap. 2) must be of high quality and have a protein content of at least 12 %. The ash content will not affect the cooking quality of the pasta to any decisive extent. Rather, high ash content implies elevated protein content, which is of advantage for the cooking properties. However, higher ash values will affect the color of the pasta, which may become discolored to the point of becoming dark brown. Crude fibers and specks show up as dark inclusions in the end products and indicate inadequate cleaning in the grinding system. The color of the semolina should be amber.

It is an advantage if the semolina particles have the most narrow particle size distribution possible, that does not exceed 0.350 mm. Fine particles absorb water more readily than coarse particles, owing to their larger specific surface area. Therefore, the more homogeneous the particle size, the more uniform their wetting. This results in a better dough, shorter mixing times and thereby smaller mixers. This was extensively demonstrated by Manser [3] and Baltensberger [4]. The mixing times as a function of the particle size are specified in Table 2.

Table 2. Mixing time as function of particle size of semolina

630–125 µm	15 min mixing time
525–125 µm	15 min
525–0 µm	15 min
350–125 µm	10 min
250–0 µm	5 min

Table 2 shows clearly that the better water absorption by fine semolina (<350 µm) enables the mixing time to be reduced.

3.3.1.2. Water and Additive Addition

The water must have a clear appearance and be neutral in terms of odors and taste. It should contain a minimum amount of minerals and must always be of drinking water grade. Foreign substances and the microbial count are typically subject to limits. The feed water temperature must be adjusted to the semolina temperature in order to obtain a certain dough temperature. The water temperature should not exceed 50°C in order to prevent protein damage. During production, a uniform water temperature must be ensured.

Since also the egg content must be guaranteed when eggs are added to the pasta, legal regulations must be observed depending on the specific country of production. The egg quantity will typically vary between one and seven eggs per kilogram of semolina. Here, too, special attention must be paid to the bacteriological criteria. High priority must be given to special processing and the use of pasteurized egg products, because this raw material is highly perishable.

To color pasta, other additives such as tomatoes or spinach can be added. They can be included in the form of powders or liquids.

3.3.2. Material Handling

The raw materials supplied from the mill (semolina, flour) are stored in bins. From there, they are fed to the extrusion press for processing. The storage system also includes equipment for regrinding material obtained during cutting of the long goods (regrind). This regrind is filled through a manual intake hopper into the regrind bin. It is first ground and then added at low rates to the semolina. As this material has been processed before it is very compact. Regrind should not be added to long goods products, if possible, and to short goods only at low ratios up to maximum 10 %. The particle size of the regrind should not exceed 0.300 mm.

If egg pasta is to be produced, an egg conditioning and feed system must be provided.

Preferably, it should be installed in a separate section of the building and be kept absolutely clean.

3.3.3. Short Goods, Long Goods, and Process Technology

Some 250 different pasta product shapes are available in the market. Pasta is classified into four main product types: short goods, long goods, nidi, and twisted goods.

Short Goods. Short goods such as elbows, spirals, noodles, etc., are extruded through a circular die and cut to the desired length by a rotary knife as they leave the die. The products first pass through a shaking predrier in order to obtain superficial drying. This is needed to prevent them from sticking together and to stabilize their shape. As a predrier, a belt-type or drum-type unit can be used, depending on the shape produced. The advantage of the drum drier is that it continuously reshuffles the product. Piled on metal decks, the short goods pass through the belt drier in about 3 or 4 h. The completely dried products can be stored in bulk storage bins or belt stackers, depending on their particular shape. Short goods lines can also be applied for making cut-out and soup pasta. Soup pasta is dried to a final moisture of 6 %. Another possibility is to produce instant pasta using an additional belt steamer ahead of the drying process.

Long Goods. Long goods such as spaghetti, macaroni, etc., are extruded through a rectangular, long die. The extruded strands are automatically spread on sticks and cut to the same length. The loaded sticks pass through the drier (approx. 5 to 6 h) and are held in intermediate storage in a stacker before packaging. For packaging, the dried product is stripped from the sticks and cut to the desired length. In certain cases, the product is cut right after drying/cooling and placed in intermediate storage in a belt stacker.

Nidi and Twisted Goods. Nidi (nests of long good pasta) or twisted goods are extruded through a rectangular, long die. The extruded strands are automatically shaped into nidi or twisted forms and deposited on frames, which then pass through the drying line. After drying, the product slides from the frames and can be storage in a belt stacker.

Process Technology. The extrusion press is the first element in a series of machines that are needed for the actual production of pasta. In the extrusion press, the various raw materials are blended, mixed into homogeneous dough and kneaded. The dough is extruded into the various product shapes and cut to length. Besides the traditional pasta extrusion press, Bühler has developed the Polymatik extrusion technology (see Fig. 1) [5], an enclosed system which has allowed the process time to be substantially reduced compared with the traditional extrusion process. The dough preparation time, for example, was shortened from the original rough 15 min, that a traditional extrusion press require,s to about 20 s. The Polymatik system also offers crucial benefits in terms of hygiene and flexibility for product change.

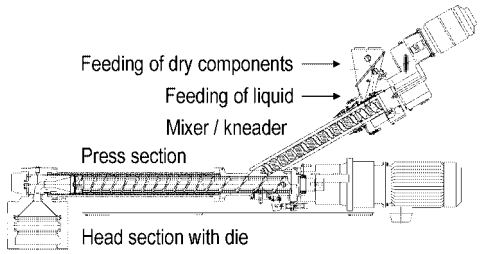

Figure 1. Polymatik extrusion press from Bühler [5]

The Polymatik press is made up of three parts: mixer/kneader, press section and head section. The mixer/kneader consists of two corotating screws, with specific kneading elements additionally supporting dough production.

Inside the mixer/kneader and the extrusion barrel shear and thrust forces act upon the dough, kneading and compressing it. An optimized kneading process in which time plays a very important part is crucial for creating the gluten framework. The conveying action of the screw inside the extrusion barrel creates the pressure needed for shaping the various products. This pressure is between 95 and 120 bar. The heat generated by dough working inside the mixer/kneader and the extrusion press is dissipated through a barrel- and head-cooling system. The dough temperature throughout the extrusion

process should not exceed 50°C. Otherwise the gluten network would be damaged, which deteriorates the cooking characteristics of the finished pasta.

In practice, attempts are often made to intensify kneading of the dough in order to offset an irregular particle size and therefore the irregular water absorption of the product in order to prevent white specks in the finished product. But the resulting high heating of the dough will at least partially destroy the protein framework, degrading the cooking characteristics of the end product.

Air entrapped inside the pasta will affect the drying process and the appearance of the end product. Therefore, the air in the dough is evacuated in the extrusion zone using a vacuum system.

In the extrusion press head, the dough is spread across the entire face of the die. The dough is then extruded under high pressure into the desired shapes through the holes of this die. For cut-out goods, a dough belt is extruded which is calibrated and then transferred to a special cut-out machine. There, specialty products such as bowties are cut out of the belt, folded and fed to the drying line.

3.3.4. Drying

As mentioned above, drying may take place in different temperature ranges. A high-temperature phase during the drying process will reduce the drying times. This, in turn, results in smaller drying lines or higher production capacities for a given space. In addition, the high temperatures guarantee the microbiological safety of the products. However, the greatest benefit of higher drying temperatures is the improved product quality. The influence of the high-temperature phase strengthens the protein framework of the product. This produces a higher cooking tolerance, lower or complete absence of surface sliminess and better chewing characteristics ("al dente") of the end product. The influence of a high-temperature phase on the product characteristics was investigated in detail in [6]. It was found that the application of high drying temperatures together with a low product moisture (<15 %) will produce the most effective results. However, these investigations were conducted on a laboratory scale. In addition, high-temperature drying improves the product quality when lower-cost soft wheat flours are added, which tend to have a weak protein framework. But it should be noted that it is impossible to use this drying process to make high-quality products from low-quality raw materials.

Whereas conventional high-temperature drying uses temperatures up to about 75°C, in the Bühler drying process "Turbothermatik" temperatures in excess of 80°C are applied.

A typical drying diagram for long goods as applied today is shown in Figure 2. It must be noted, however, that the values specified are merely approximate, since the optimal drying curve has to be determined first for every single product, varying with the shape, raw materials, line design, etc.

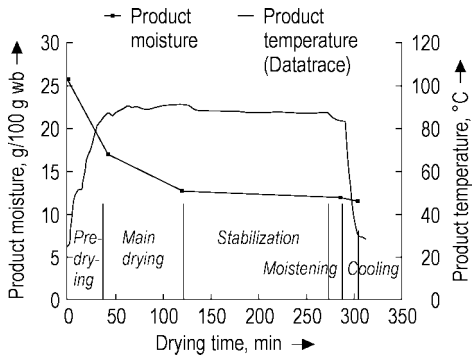

Figure 2. Example of drying diagram for long goods [7]

To enable optimal control of the drying process, the drier is of modular design and is divided into zones allowing individual climate control (Fig. 3). In zones A1 to A4/A5, predrying takes place. B1 is the main drying zone, and C1 the stabilization zone. In C2, moisture is added to the product, and in D1 and D2 the product is cooled to ambient temperature. Each of these zones is equipped with a climate probe that measures the temperature and humidity of the drying air and transmits these values to the control system. The system then controls the heating valve and the fresh and exhaust air system of the drier to ensure a consistent climate. At the start and end of production, it is possible to inject steam in order to obtain the required air humidity. This required humidity is then achieved through the

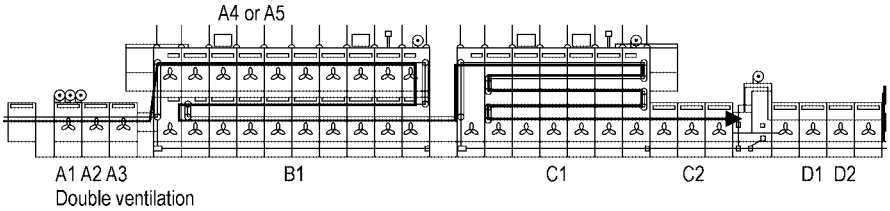

Figure 3. Schematic representation of the C-line long-goods drier

product moisture released when the drier is fully loaded.

In the case of long goods, for example, the product is dried in the predrying zone within a very short time from approximately 30 % to about 16 % product moisture. The process parameters must be adjusted as a function of the raw material quality. The settings must be selected such that no crusting of the product is created. Otherwise, it will not be possible for the moisture remaining inside the product to be adequately released. For thermodynamic reasons, this entrapped moisture will create stresses in the product, which may lead to product breakage during storage. During the main drying time, the product is ideally dried to a final moisture content of about 12 to 12.5 %.

The longest drying phase in terms of time is the stabilization phase. It takes place at about 2 to 4°C lower temperatures than the high-temperature phase (main drying). The main purpose of stabilization is to relieve the stresses inside the product and to prevent any further drying of the product.

Providing the drying diagrams are properly applied, it is also possible to improve the color of the pasta in high-temperature drying. On the one hand an enzyme inactivation takes place, and on the other hand the creation of color pigments (Maillard reaction) can be reduced if, with low product moisture, the drying temperature for stabilization is reduced again.

At the end of the stabilization phase, the relative air humidity is increased again to add moisture to the product and to balance the moisture gradient in the product. In the subsequent cooling zone the excess heat is extracted from the product in order to stop desorption. This prevents further drying of the product and reduces the product temperature to the ambient temperature.

For lower capacities, a single-housing drier (see Fig. 4) will suffice, where the entire drying process takes place in the first housing. The main drying zone B1 is then located in the upper segment, and stabilization (C1) is carried out in the bottom section (B2 in Fig. 4).

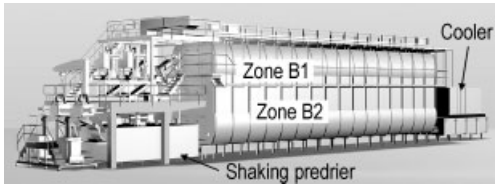

Figure 4. Bühler C-line short goods line with single housing drier

3.3.5. Intermediate Storage

To allow efficient and economic operation, today's production lines are in service around the clock. They are only stopped for maintenance and cleaning. On the other hand, packaging of the finished products is typically done in one- or two-shift operation. This, of course, requires a possibility of storing the products while the packaging line is shut down. Depending on the specific product shapes and the line capacity, various options are available.

Short Goods Bins. For so-called rolling short good shapes such as bows, rigatoni, penne, etc., storage bins are excellently suited and likely the most economic solution. The products are transferred from the cooler via a bucket elevator and belt conveyors to the storage bins and are retrieved from storage again by belt conveyors, which move them to the packaging lines. Care must be taken to ensure that no bulky products are stored in the bins, because the resulting high weight would cause the products to break under their own load. Moreover, bulky products have

poor flow characteristics, affecting discharge from the storage bins.

Short Goods Belt Stackers. For storage of bulky products such as noodles, belt stackers are used. As a result of the very shallow depth of product on the individual belts of the stacker, the weight of the product is low so that degradation of the products is prevented.

Long Goods Stacker. After discharge from the cooler, the loaded sticks are automatically transferred to the stacker. The product is held in intermediate storage on several tiers, hanging from the sticks. When the packaging line is started, this also automatically activates the stripping and cutting machine at the outlet of the stacker. There, the product is stripped from the sticks, cut to the proper length and transferred to the packaging line via cascades. The trimmed-off bends and ends are conveyed pneumatically to the regrinding system described above. The empty sticks are returned to the spreader by a stick return system, which is installed underneath the drier elements.

Nidi and Twisted Goods Stacker. In the same way as with the short goods stacker, the product here is stored on belts. But as nidi and twisted goods require very gentle treatment, bucket conveyors are used instead of elevators and belt conveyors for feed and discharge.

3.3.6. Thermal and Electrical Infrastructures

The energy consumption of state-of-the-art pasta production lines equals that of traditional production systems. The energy costs (oil, electric power, cooling, steam) account for about 2 % of the total production cost. In view of the high cost of energy, the manufacturers of pasta production equipment are forced to give consideration also to this point. They must therefore manufacture machinery operating with the highest possible economy and minimum energy losses.

Heating. For heating the driers, hot water is used. Depending on the required drying temperature, the water must have an inlet temperature as high as 130°C. In many cases, this will require the use of a pressure system. The most proven design for a heating system is one equipped with a primary and a secondary circuit for each drier. This enables very accurate and fast temperature control.

Steam as a heating medium for driers has proven inadequate, since steam is too sluggish to allow fast temperature changes. If a steam heating system is installed, hot water is generated in heat exchangers.

The size of the heating boiler depends entirely on the line capacity. As an approximate value, a heating capacity of about 24 000 kJ/h can be assumed, based on a pasta output of 100 kg/h.

Electric Power. For the total power supply of all motors of a production line including the extrusion press, about 11–13 kW must be assumed, based on a pasta output of 100 kg/h. Here, too, the actual consumption values are about 20–25 % lower.

Power consumption for the bin storage system and the packaging equipment depends on the design of these installations and must be defined on a case-by-case basis.

Cooling. For the extrusion press barrel and the cooler after the drying line, cold water is used as a cooling medium. Per 100 kg/h line capacity, about 10 500 kJ/h must be assumed.

Steam. Steam is used only for steam injection during the starting and stopping phases. The most suitable type of steam is saturated steam of max. 1 bar pressure.

Compressed Air. To actuate the valves of the bin storage system and the packaging equipment plus those on the production line, a compressed air system is installed.

3.3.7. Pasta Quality

High-quality pasta products are smooth, hard and to a certain extent resilient or elastic. They should be transparent and have a rich yellow color (depending on the raw material used).

Long goods, especially spaghetti, must withstand considerable flexing stresses without breaking. Fracture surfaces must have a neat

and glassy appearance. Of course, the quality of pasta becomes apparent when consumed.

Pasta swells during cooking and absorbs water. Good products absorb up to double their own weight of water or even more. Their volume will increase three- or four-fold.

Good pasta retains its shape and has a certain "bite" (al dente characteristics).

If properly stored, pasta has a shelf life of several years.

4. Modified Flours

4.1. Definition and Possible Applications

Modified flours are thermally or mechanically treated cereal products, which have many different applications in food industries. Therefore, those products have to meet special, adapted, and functional requirements. This asks for the use of an extremely flexible process technology for the production of modified flours. The requirements can be achieved with the help of cooking extrusion technology such as HTST (high temperature, short time) treatment. Table 3 lists different products and the corresponding use of modified flour.

Table 3. Applications of modified flours in food industries

Baking industry	Binder for water, prolongation of freshness
Sauces	binder, thickening agent
Soups	thickening agent
Sausages	binder, binder for water
Meat, fish	filler, adhesive agent for breadcrumbs, breadcrumbs
Tablets	substrate, replaces modified starches
Deserts	binder for water, thickener
Baby food	basis, carrier for calories, thickener

A thermal or mechanical treatment means that flour, together with water, under application of heat and shearing, is kept for a certain time above the gelatinization temperature of starches. During this time, defined physical changes of the functional properties of the product, like viscosity behavior, water absorption index and water solubility index or cold-paste viscosity, occur.

4.2. Production

4.2.1. Process

A simple thermal treatment of cereal flours can be carried out industrially with the help of autoclaves, jet cookers with spray driers or drum driers, flour heat treatment, etc. However, using cooking extrusion, a variable thermal and mechanical treatment can be achieved.

The most important criteria for the desired product properties and qualities are the process and the setting of parameters of temperature and pressure. In cooking extrusion technology, the introduction of energy into the matter can be achieved by feeding steam into the product above or directly into the extruder. This can be done by convection via the barrels or by energy dissipation via shearing. Heating up and preswelling of flours in preconditioners is nowadays widely applied. By varying the ratio between thermal and mechanical energy introduction, a direct influence on structural changes, e.g., the molecular decomposition of starch, can be taken.

Thus, the functional properties of modified flours can be influenced systematically. The direct characteristic correlation of the structure is described in [8].

Furthermore, extrusion is distinguished by its continuous operation at short term and high temperatures and pressures, resulting in lower production costs compared to traditional processes. This is mainly caused by lower investment costs at reduced space and human resource requirements.

In addition, the process is very flexible and easy to adapt to market needs concerning throughput and product quality together with a simple controllability.

4.2.2. Production Plant for Modified Flours

Extrusion plants for the production of modified flours are already proven in industrial application in throughput ranges from 200 kg/h up to 4000 kg/h. The plants include the following process steps: premixing, dosing/feeding, preconditioning, extruding, drying, grinding, and sifting and, if necessary, mixing/blending and packaging.

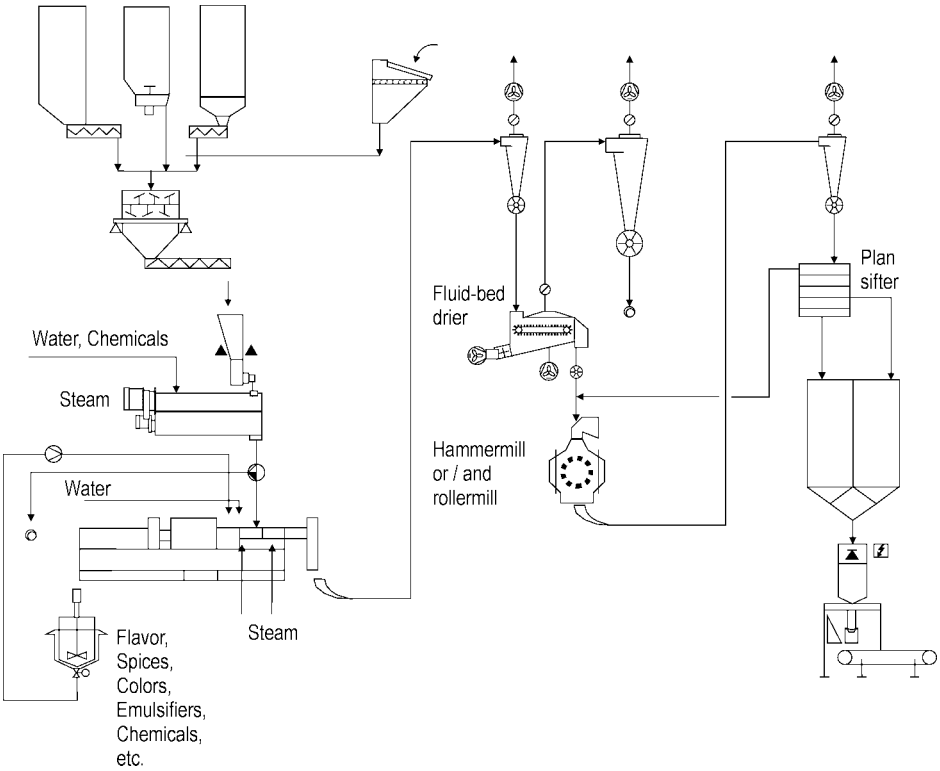

Figure 5. Flowsheet of a plant for the production of modified flours (Source: Bühler) [9]

The flours can be premixed with other ingredients and are then fed by continuous dosing into the preconditioner, where they are mixed intensively with water and steam to achieve a hydrothermal pretreatment of the blend. Due to the thermal modification quite high degrees of gelatinization of the starch can be achieved already. Thus, if sufficient water is added, the high amount of starch contained, e.g., in wheat flour can be gelatinized to more then 95 % (as determined by differential scanning calorimetry, DSC).

The following extruder mixes, shears, cooks, and plastifies a dough and afterwards presses it through a die. In consequence of the resulting difference in pressure from inside the die to atmospheric conditions, the product expands and partly dries by an abrupt evaporation of the steam. The extrudates are then transported pneumatically into a drier. Depending on the desired water content of the finished product, the extrudates will be dried and subsequently ground. After milling, the material is sifted to classify it into the desired granulation ranges. Figure 5 shows an example of a flowsheet of a plant for the production of modified flours.

4.3. Technological Aspects of Extrusion Cooking

Under the conditions of extrusion cooking, depending on the process variables, the functional properties of the flours are significantly changed. The reactions occurring inside the extruder are multidimensional and depend on a wide range of machine- and raw material-specific process parameters. Due to these multidimensional reactions, a differentiation of the influence of the single parameters on the product properties is hardly possible.

Therefore, an analytic model of the system was developed. It partitions the process in partial stages and differentiates between material properties, variable and fixed process parameters, system parameters, product changes, and product properties.

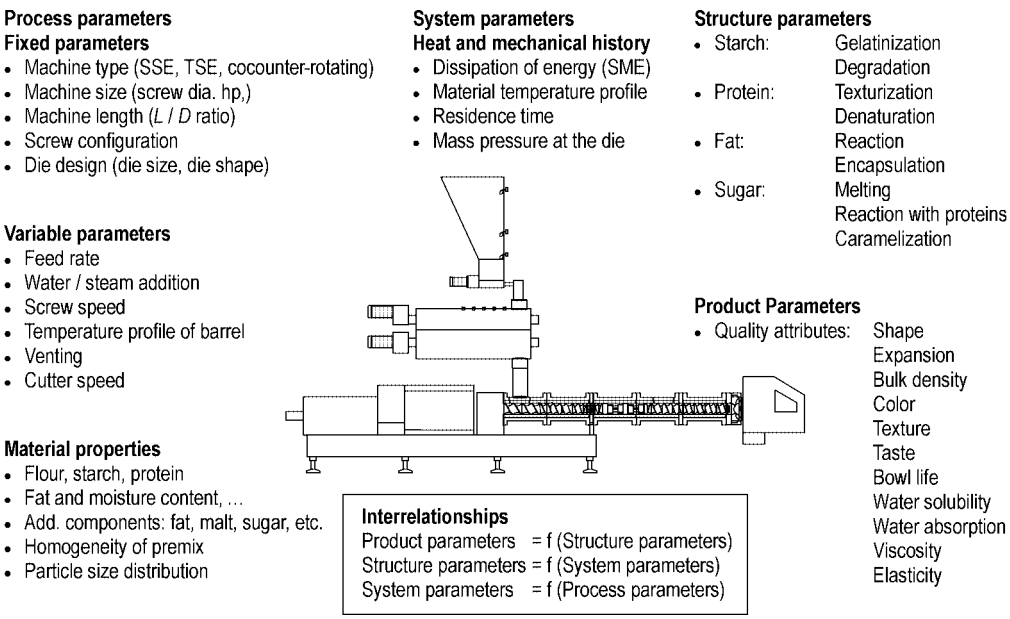

Figure 6. Overview over the unit operations of the cooking extrusion process (Bühler)

A systematic approach allows to specifically differentiate the influences of the extrusion parameters on the properties of the modified flours and allows further to use the knowledge about these dependencies.

Figure 6 depicts the extrusion process on the basis of an analytic model and gives an overview on the possible unit operations.

To understand this model, it is important to know which processes take place inside the extruder and how they occur.

An exemplary course of temperature and filling degree over the whole length of the extrusion process is shown in Figure 7. From that a specific time-temperature behavior for a certain point of process results.

Inside the preconditioner, flours are intensively mixed with 5–25 % water and up to 7 % steam. Thus, each particle of flour is heated up immediately with the help of condensation heat and on its surface the water is equally distributed to achieve an easy, quick, and complete penetration.

In this way the swelling of the starch granules up to high degrees of pregelatinization is induced. The pregelatinized flour is then directly fed into the extruder. There, it is also possible to feed additional water or direct steam to exactly control the introduction of energy into the matter to reach the desired product properties.

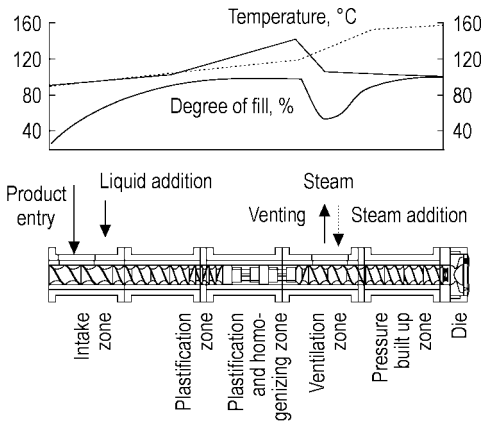

Figure 7. Schematic illustration of profile lines inside the cooking extruder

By possible feed of additional water into the mixing zone the bulk material is transferred into a dough. The matter is then transported into the plastification zone, where it is exposed to shear forces by increasing the filling degree, the pressure and under influence of the speed of the screws. Thus a dissipation of energy is caused which partly is transformed into heat.

This causes a further decomposition of the material up to a complete decomposition and plastification of the starch granules. The causative parameters for this modification are the water content, the temperature, the introduction of mechanical energy and the pressure. Inside the extruder it is possible to consecutively place multiple plastification and tension release zones, which enables to stepwise shearing and releasing the product to specifically set some temperature peaks. Thus, damaging of the material can be controlled exactly.

After the mixing and shearing zone, the material passes the dwelling or reaction zone at increased temperature. In the following zone directly before the die pressure is built up and additional energy is introduced. Filling degree and dissipation of energy are increasing. Thus, the temperature rises again. Before the die, the dough is plastified, whereas the contained water is finely dispersed. As a result of the high pressure of approx. 20–150 bar and temperatures between 120 and 180°C the water is not in the vapor, but still in the liquid state. Forced by the screw conveying action of the extruder the cooked mass is pressed through the die and then immediately cut by a rotating knife. Under the abruptly upcoming atmospheric conditions, the water contained in the dough is promptly evaporized and blows up the elastic dough. Caused by the finely dispersed droplets of water, each droplet results in a bubble and a foamy structure occurs. The expanded vapor exhausts, thus the whole process is named a drying by expansion. Via the aforementioned parameter, the degree of expansion can be controlled.

The evaporation of about 5–10 % of water effects—by removal of the evaporation heat—a quick cooling and hardening. The generated foamy structure will persist.

4.4. Effects of the Process Variables on Parameters and Functional Properties of Modified Flours

As a result of trials performed by the Bühler company and other published analyses [8, 10, 11] the following relations and functional dependencies could be hypothesized.

4.4.1. Effect of Process Variables on Parameters

The cold-paste viscosity can clearly be described in dependence of the water content at given machine and process parameters by the system parameters specific mechanic energy introduction and specific thermal energy introduction. Caused by the available water in the dough, the viscosity of the dough in the extruder and hence the resistance of the screws inside the extruder is influenced by the product, whereby shear stress and dissipation of energy are predetermined. Depending on the water content, at otherwise constant conditions, a wide range of qualities can be produced

- from almost burned product, at highest energy introduction and very high temperatures inside the dough,
- over dextrinized flour, thus highly decomposed products with lowest cold-paste viscosity, and
- maximal gelatinized starch with highest cold-paste viscosity,
- up to partly cooked and uncooked products at very high water contents and very low introduction of energy.

Dependencies of the Process Variables. Screw speed and water content have influence on specific mechanical energy (SME) introduced in the product. A rising screw speed causes increasing of the shear rate. This means higher SME and thus the temperature of the material rises. With low water content, the viscosity of the dough inside the process section of the extruder is much higher and thus the introduction of energy by increased shear stress rises significantly, resulting in a higher end temperature of the product. Along with the introduction of energy, caused by the degradation of ingredients a decrease in viscosity results and thus the pressure before the dies does not rise accordingly.

4.4.2. Effect of Parameters on Functional Properties

The functional properties of the finished products such as indices of water absorption and

water solubility and cold-paste viscosity depend strongly on the conditions prevailing inside the extruder, i.e., on the parameters, specific mechanical energy input, resulting temperature, and retention time.

The water absorption increases enormously with decreasing energy input and decreases slightly with increasing temperature of the material. The water solubility increases with increasing energy input, i.e., with increasing SME or also temperature. The cold-paste viscosity increases strongly with decreasing specific energy input. All three functional properties depend also on the water content.

4.4.3. Effect of Functional Properties on Finished Products

Definitions for the definite impacts of those parameters on the finished products can only be generalized to a very limited extent, because all producers of finished products containing modified flours have their specific recipes, whose specific composition have to fulfill the requirements of the corresponding markets. However, the recipes will not be published by the producers.

An exception are the publications of the German Federal Research Institute for Cereals [12, 13], in which the behavior of viscosity and the impact of extruded wheat flours on the quality properties of pastries and wheat bread are discussed. It was detected that addition of 5–20 % of modified flour to filo dough caused desired changes in the rheology of the dough. The pleasant impression of short chewing and the freshness were improved. An addition of 5 % of modified flour to a wheat bread dough resulted in an increased dough and bread yield at higher water binding capacity. Also here, the freshness was improved.

Physical alterations of flours to reach specific product properties become more and more important. This is caused by increased demands on the produced food as well as by the increasing necessity to declare all additives, as requested by EU directives. Due to the increasing restrictions regarding the authorization of chemical additives for flours, more and more alternative solutions have to be found and used to reach the same effect as with chemical agents. The treatment of flours by HTST extrusion enables production of a wide range of those products, whereas in general no additives have to be used, because modified flours are simply cooked and sheared under pressure with the only addition of water.

5. Breakfast Cereals

5.1. Definition and Historical Background

End of 19th century ready to eat (RTE) breakfast cereals were among the first deliberately conceived "health foods," developed in Battle Creek, Michigan, USA to make more palatable the diets based on whole grains that were advocated then. Some of the processes and products developed in the early days of the cereals industry are still in use.

Early products included flakes made from corn grits and wheat berries, shredded wheat biscuits, and nuggets of barley and other ingredients baked in loaves and then milled and dried. Today, RTE cereals are profitable and diverse segments of major food companies. A big player in the market collectively is the category of private-label cereals made by several small- to medium-sized firms.

The private-label products are usually imitations of the branded products, but sell for less because they have lower marketing costs, the companies do not invest as much in research, and sometimes use less expensive processes.

5.2. Processes

The major cereal processes are: cooking, flaking, gun puffing, shredding, and extrusion.

5.2.1. Batch Cooking

As described in [14] the early cereal production included separate mixing, cooking, forming, drying and coating steps. Typically, many of these process steps were, and often still are, batch operations. For example, whole wheat berries or corn grits are cooked in pressure vessels with added malt, salt, and other ingredients

using direct steam injection. The objective is to gelatinize the starch and temper the grains with enough moisture to make them plastic without turning them to soft mush.

After cooking, the grains are cooled and slightly dried before forming by passing between polished steel rolls to form flakes. Whole grain rice is also cooked and then lightly "bumped" between rolls before forming by toasting in an oven.

In the traditional flake process, flakes from the rolls are dried and toasted in conveyor ovens. Some modern ovens use high velocity jets to fluidize the bed of flakes and shorten the drying time.

If flakes have to be coated with sugar or flavors, they pass through a coating drum or belt and then are dried again. Coated or uncoated flakes often have a vitamin emulsion sprayed on them as a final step, followed by another drying step and then packaging. The vitamins and minerals, if used, are added last because they would not survive the high temperature treatments or earlier process steps.

Breakfast cereals are seen as desirable vehicles for delivery of added nutrients because they can represent a significant portion of a daily diet, are usually eaten with milk, and have wide appeal. Often, nutrient fortification is an important part of a given brand's image and marketing message.

5.2.2. Gun Puffing

Gun puffing starts with cereal flour or whole grains, depending on the products. Some producer make puffed whole wheat and rice, while others use oat flour and other ingredients to make some expanded products. Flour-based products are formed using low-pressure extruders, similar to those used for pasta. The pellets are dried and then heated in a closed pressure vessel. The pellets are released from the elevated pressure either by quickly opening a closure or by transfer through a special valve. The sudden pressure change in the softened pellet causes expansion of water vapor and quick cooling, so that the pellet expands and then hardens. The porous shape retains a crunchy texture even in milk.

Puffed pieces are dried further, coated and enriched as mentioned earlier, and packaged.

5.2.3. Shredding

Shredding involves cooking whole grains to a precise moisture content and then passing them through batteries of special rolls in which one roll is grooved and presses against a smooth roll. High pressures are developed as the grain is forced into the grooves, so the rolls tend to be short to control deflection. In a commercial line, there are many banks of rolls. The threads of wheat are laid over one another in mats and cut into biscuits. A variation is to have cross grooves so that threads are attached to one another.

In some processes, cooked grains may be stored in a relatively moist state before forming. It may be that this step contributes to final flavor by permitting chemical and biochemical reactions to occur. In general, it is observed that the traditional, batch, relatively slow processes seem to produce the more complex flavored products, though there are constant efforts to increase speed and efficiency.

5.2.4. Cooking Extrusion

One of the great improvements in cereal processing was the application of cooking extrusion. An extruder can combine in one machine the operations of mixing, cooking, forming, and expansion. At the time when cooking extrusion began, no extruders existed that were intended for foods, thus cereal producers used equipment from the plastics industry. Today, food extruders are produced by many manufacturers. For a more detailed description of the technological aspects of extrusion cooking, see Section 4.3.

Because flours rather than intact berries or grits are generally used for production of extruded cereal products, and because HTST processes are employed, it is a great challenge to match with extrusion technology the flavor and texture of more traditionally made products. The goal of the branded cereal makers is to deliver good taste in combination with healthy nutrition. Lately, this has meant a focus on products from whole grains with high bran or fiber content. Fibers are essentially inert and interfere with expansion. Some products can be made by alternative processes, such as gun puffing or extrusion.

Flakes can be made by extruding pellets and then flaking those pellets. This permits use of flour and incorporation of other ingredients in comparison with using whole berries or grains.

5.2.5. Coating and Inclusions

Sugar coating is a complex topic in its own way. Coatings may be transparent or opaque and may include artificial sweeteners such as aspartame and noncariogenic materials such as sugar alcohols. High-solids coating solutions must be kept hot and are subject to discoloration. Application can be challenging because the solutions can crystallize and harden quickly.

Fruits, such as raisins, have been added to cereals for many years, but in the search for new products, other inclusions, such as nuts, freeze-dried fruit, yogurt and granola (a cereal mix itself) have been added. These create processing challenges in feeding at correct rates, prevention of segregation, matching moisture activity, and preventing creation of fines.

The water activities of raisins and cereal flakes are quite different, for instance, so that when enclosed in a sealed package, the flakes can become soggy and the raisins dried out. Early packaging used waxed paper in cardboard boxes, which actually permitted some moisture to escape. When cereal packaging was changed to polymer liners, which are better moisture barriers, the equilibration issue became more noticeable. One solution has been to add solids to the raisins, such as sugar or cereal fines coatings, to reduce water activity.

Fines are an almost inevitable byproduct of cereal manufacture and have become a useful ingredient in other products, such as breadings and coatings.

Ready to eat cereals are an interesting food category not only for the variety and challenges of the many processes, but also for the several ways in which they can be analyzed. There are cereals for adults and for children; sweetened and unsweetened; short-lived novelties and 100-year-old standards; originals and copies; whole grain and flour-based; branded and private label. For manufacturers, RTE cereal has been profitable, challenging, and a sources of technical accomplishment. For consumers, it continues to be nutritious, enjoyable, and generally a good value.

5.2.6. Production Plant for Breakfast Cereals

Extrusion plants for the production of breakfast cereals are for a long time proven in industrial applications in throughput ranges from 200 kg/h up to 2500 kg/h. Figure 8 gives an example of a production plant.

Breakfast cereal plants can include the following production steps: premixing, feeding, preconditioning, extruding, predrying, flaking, toasting / cooling, coating and final drying.

The premixed raw materials are fed to the preconditioner, where they are mixed intensively with water and steam to achieve a hydrothermal pretreatment. A high degree of pregelatinization can be achieved already. Inside the extruder the material is further mixed, sheared, cooked and plastified. The dough is then pressed through the extruder die.

In case of the production of *direct expanded cereals*, the temperature at the die has to be higher then 100°C and the pressure must be higher than product elasticity and atmospheric pressure. In consequence of the hence resulting difference in pressure and temperature from inside the die to atmospheric conditions, the product will expand and partly dry by an abrupt evaporation of the steam. The moisture content of direct expanded products after being cut at the die is about 8–12 %. The extruded products is then transported to the drier and afterwards eventually to a coating drum to apply, e.g., a sugar coating and then to the final drier. The finished product will have a moisture content of below 3–4 %.

In case of *indirect expanded products*, e.g., corn flakes, the expansion at the die should be prevented. Therefore, the temperature of the material before the die has to be lower than 100°C and the pressure should not exceed the sum of product elasticity and ambient pressure. To achieve this, at the end of the process section of the extruder degassing for the evaporation of steam and cooling has to take place.

The extruded pellets have a moisture content of about 20–22 % and are pneumatically transported to an intermediate drier. Here, mainly the

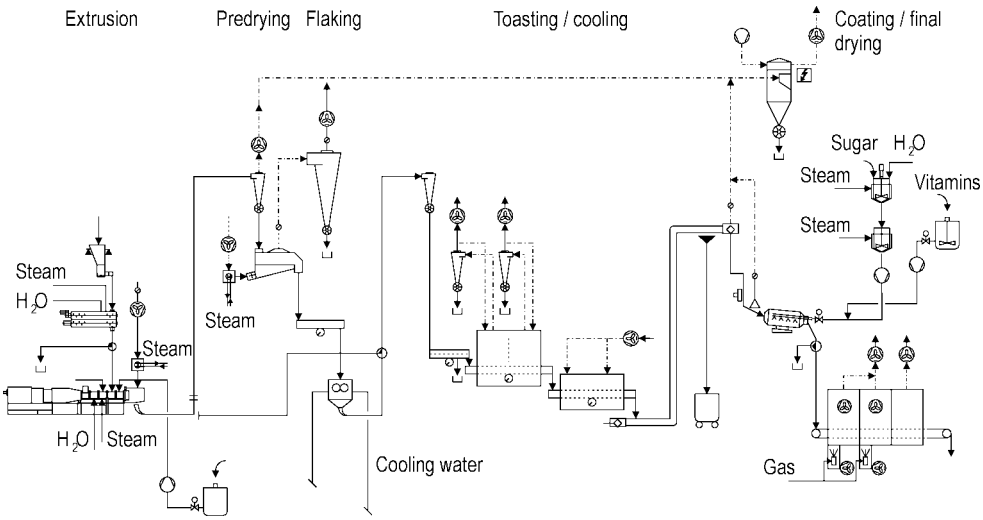

Figure 8. Example of a production plant for breakfast cereals based on extrusion (source: Bühler)

surface is dried to prevent the pellets sticking together and an equalization of moisture distribution inside the pellet takes place. After the intermediate drying step, the pellets are flaked with about 16–18 % water. The expansion of the corn flakes and thus the development of the bubbly crunchy texture of corn flakes take place in the toaster. Temperatures of above 250°C lead to an immediate evaporation of part of the moisture content inside the flakes and thus to an expansion. The finished corn flakes can then optionally be sugar-coated and dried.

Corn flakes also have a moisture content of below 3–4 %.

6. Reconstituted and Fortified Cereal Grains, e.g., Rice Kernels

One of the most fundamental decisions underlying food fortification schemes is selecting appropriate foods to be fortified with the essential micronutrients lacking in a population's diet. Criteria to identify potential food fortification vehicles generally include selecting a food that is commonly eaten by the target groups, is affordable and available all year long, and is processed in such a manner that fortification is technically feasible and can be done economically. Staple foods such as wheat flour and sugar have been popular foods to fortify in order to address micronutrient deficiencies in several developing countries.

6.1. Importance of Rice as a Food

It has been estimated by Juliano [15] that half the world's population subsists wholly or partially on rice. Globally, rice accounts for 22 % of total energy intake [16]. Ninety percent of the world crop is grown and consumed in Asia. American consumption, although increasing, is still only about 11 kg per person annually, as compared with 90–181 kg per person in parts of Asia [1]. Rice is the only major cereal crop that is primarily consumed by humans directly as harvested, and only wheat and corn are produced in comparable quantity.

The trends in rice consumption are closely linked to rice production in a number of rice-producing countries as illustrated in Figure 9.

Low energy and protein intakes are common nutritional problems for people in rice consuming countries. The FAO estimates that about half the population of South Asia do not have adequate energy intakes. Nutrition indicators compiled for 34 rice-consuming countries indicate that the incidence of low birth weight, infant mortality, mortality of children fewer than five years of age, and prevalence of underweight children are considerably higher in these countries than in other countries [17].

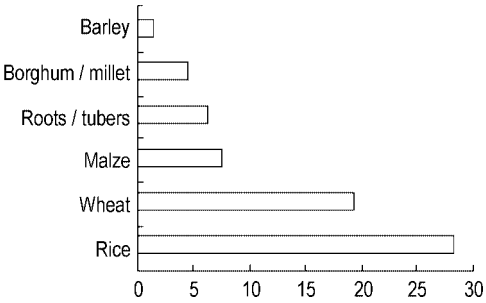

Figure 9. Percent calories contributed by stable foods to diets–all developing countries (Source: FAO/ESS (averages 1992-1994))

6.2. Rice Milling and Nutrient Losses

Milling of rice is different from that of other cereals, because the objective is to produce a maximum yield of unbroken milled grains rather than a flour or meal as in the case of most other cereal grains. Processing the unhulled rice grain (also called paddy or rough rice) involves cleaning, milling to remove the hull, germ, and bran layers, and sizing to produce white uncoated rice. A coating of talc and glucose may be added to improve appearance; however, this is not permitted in some countries. Milled or white rice represents 40–76 % extraction of rough rice. The B vitamins and iron are found primarily in the germ and bran layers, and are therefore removed in the milling process. It has been estimated that in the course of milling brown rice to white rice approximately 80 % of the thiamine is removed [18]. Other nutrients contained in the bran layer are also lost, including niacin, iron, and riboflavin.

Nutrients can be preserved in the rice grain by parboiling (see → Cereals, Chap. 7.6). Parboiling transfers the nutrients contained in the bran layer to the inner endosperm layer prior to milling and removal of the bran. The parboiling process involves soaking of rough rice and applying heat and then drying and milling. Parboiled rice is produced in India, Bangladesh, Burma, Thailand, Sri Lanka, and other Asian countries by both traditional and modern parboiling processes [19]. Parboiling may require expensive equipment and can result in golden color rice that is often not acceptable to consumers. Converted rice was developed with a similar process but uses pressure to transfer nutrients to the inner endosperm layer. Enrichment methods have been developed to add synthetic vitamins and minerals to rice to replace those lost in milling. Primarily iron, thiamine, niacin, and riboflavin are added, although other nutrients, including pantothenic acid, vitamin B6, calcium, and folate have also been included in some enrichment premixes. In addition, rice has been used as a vehicle for vitamin A fortification in various pilot programs.

6.3. Extrusion Process for the Production of Tailor-Made Rice Products

Using cooking extruders it is possible to produce granules, which are similar to rice kernels, out of broken rice or rice flour. By setting the optimal process parameter and recipes, those extruded rice granules can feature a wide range of formats and properties. The main features of the rice granules such as color, shape, size, and texture, ingredient composition, cooking behavior and optionally also the digestion properties can be determined and tailor-made for each application due to process and recipe settings.

Figure 10 illustrates the main process steps in the production of extruded rice granules. Figure 11 shows a possible implementation into a machinery flowsheet.

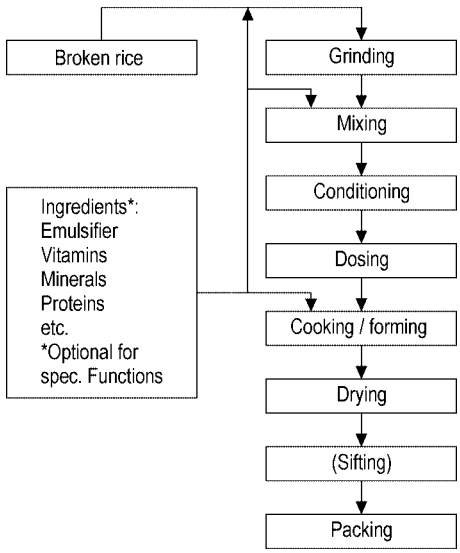

Figure 10. Main steps in the production of reconstituted rice

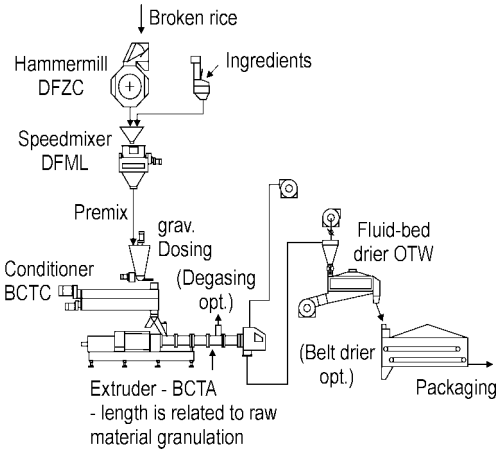

Figure 11. Flowsheet of a production line for reconstituted rice (Source:Bühler)

Due to their adaptable properties, the extruded rice granules can be used for different applications. Depending on the process parameters, the granules can be used as instant rice granules, as quick cooking rice granules with a coking time of 2–7 min or as an additive to mix into natural rice.

6.3.1. Quick-Cooking Rice

Two quick-cooking rice products exist: quick-cooking rice granules and quick-cooking rice kernels produced from natural rice kernels. *Quick-cooking rice kernels* can be produced with the help of diverse thermal processes out of dehulled, whole rice kernels. They have cooking times between 8 and 14 min. Such quick-cooking rice is only hardly to differentiate from, e.g., parboiled rice.

Quick-cooking rice granules, produced with the help of a cooking extrusion process, are used for similar applications as traditionally manufactured quick-cooking rice, i.e., often packed in transparent packing. Therefore, their appearance is highly important. A reconstituted rice granule should optically not differ from traditionally manufactured quick cooking rice. But also the texture should not be different because quick cooking rice is mostly used as a stand-alone product, e.g., as a side dish with meat. A rice product which differs too much from natural rice would not generate satisfied customers.

6.3.2. Instant Rice

Before the introduction of rehydration, all instant rice products available on the market differed substantially from the appearance of natural rice. Instant rice is produced by intensive cooking or steaming with a following drying process. During these process steps, the original shape of the rice kernels is more or less lost. Due to the rehydration with hot or even boiling water, the optical and sensory differences to natural rice will largely disappear.

Instant rice, manufactured with extruders show the same properties. The kernels show a certain expansion and a porous surface, which enables a quick absorption of the water during rehydration. The optical similarity to natural rice is in this application not very important because those kernels are mostly already mixed with other ingredients together in a menu, which has only to be doused by hot water. Thus the single kernels are not eyed as such. Here, after rehydration, “only” the texture has to be correct.

6.3.3. Additive Product

To be accepted by customers, rice granules which are added to natural rice to increase the yield, have to have a similar appearance as the other kernels. Also the texture is very important and, the larger the percentage of reconstituted rice, the more important becomes the similarity.

The specialty within this application is the cooking time. Rice granules which are produces with extruders have per se a short cooking time. When they are mixed with natural rice, their texture has still to be pleasing at the time the natural rice is cooked. This is rather significant because the ratio of reconstituted rice granules can be up to 10–20 % (refers to the fraction of broken rice during milling).

A variation of the cooking stability and thus of the texture of the rehydrated rice can be done by setting of different parameter. Grinding of rice and thereby the choice of the right granulation can influence the texture, whereas fine flours lead to a more soft and sticky texture than finer or coarser semolina. Raw materials coarser then 1000 µm can influence the optical quality because a less smooth surface of the kernel can

be generated. Besides this, the cooking stability can be influenced by additives mixed to the raw material. Substances which supply cross-linkers, such as sulfur or phosphates, can create a networking of the flour ingredients which leads to a strengthening of the matrix. Some dietary fibers or hydrocolloids can also be utilized to influence the stability of the matrix, so that during the much longer cooking time of natural kernels the reconstituted rice granules as much as possible keep their form, texture and appearance.

6.3.4. Fortification of Reconstituted Rice

A further option is the possibility to add functional ingredients during the extrusion process. Hereby, the fortification, respectively the addition of several substances to the extruded rice granules is possible. For such a fortification several conditions have to be fulfilled. The additives must be stable to temperature (short term up to 120°C) and shear, so that the desired content can be reached under economically maintainable conditions. Furthermore, the influence of the additives on the production process has to be taken into consideration. An unsuitable riboflavin formulation, e.g., can lead to a yellow colored "curry rice". In case of addition of an unsuitable iron formulation, oxidation can create a brownish product.

Basically all substances known as functional ingredients are suitable as additives. The main groups are vitamins (especially of groups A and B), minerals, and trace elements (e.g., Ca, Fe, Mg, Se,. . .), proteins or several essential amino acids, fatty acids, antioxidants as well as the so-called pro- and prebiotic ingredients, such as oligosaccharides and dietary fibers. Finally all substances can be added which are accepted by customers and whose benefit can be communicated. Thus very special products can be tailor-made for diverse markets and applications. For those regions of the world, where nutritional deficiencies caused by the nearly exclusive consumption of rice occur, fortified rice kernels can be produced which can improve the deficiencies of vitamin A and other vitamins and minerals.

An interesting form to supply a huge percentage of population with essential nutrients is the possibility to produce rice granules with a high concentration of additives, so that the recommended amount of the added vitamins and minerals can be provided by addition of those kernels to normal white rice in the ratio 1:100 or even 1:200. The production of the highly fortified kernels can be organized centrally, whereas the addition to white rice can take place in each small rice mill all over the country.

By the addition of the corresponding vitamins or minerals, products can be specified for the particular deficiencies of different population groups, e.g., for older people or children by feeding in school.

7. References

1. USDA: World Agricultural Production, United States Department of Agriculture Foreign Agricultural Service, 2005, http://www.fas.usda.gov/wap/circular/2005/05-08/tables.html, 18.07.2006
2. Unione Industriali Pasta Italiani, 2006, http://www.unipi-pasta.it/dati/TA2005.HTM, 18.07.2006
3. J. Manser: "Feinheitsgrad von Durum-Mahlerzeugnissen aus der Sicht der Teigwarenindustrie," *Getreide Mehl Brot* **39** (1985) 117–123.
4. W. Baltensperger: "Granulation von Grießen für die Teigwarenherstellung," *Tecnica molitoria* 1996, 113–119.
5. W. Seiler: Polymatik® Technology, Bühler Specialist Article 2006.
6. W. Seiler: Drying, Bühler Specialist Article 2006.
7. C. Zweifel: Influence of high-temperature drying on the structural and textural properties of durum wheat pasta, PhD Thesis ETH No. 14073, ETH Zürich, Switzerland 2001.
8. Th. Brümmer: Charakterisierung molekularer Strukturparameter kochextrudierter Maisstärken und Ermittlung von Struktur-Eigenschaftsbeziehungen, PhD Thesis TU-Berlin, D 83, FB15, Nr. 151, TU-Berlin, Germany, 1999.
9. Bühler AG: Technical documents and standart flowsheets, 2006.
10. Ch. Millauer: Untersuchung der rheologischen Eigenschaften von Weizenstärke und Weizenmehl, PhD Thesis D 83, FB15, Nr. 26, TU-Berlin, Germany, 1994.

11. B. V. Lengerich: Entwicklung und Anwendung eines rechnergestützten systemanalytischen Modells zur Extrusion von Stärken und stärkehaltigen Rohstoffen, PhD Thesis D 83, FB13, Nr. 165, TU-Berlin, Germany, 1984.
12. W. Seibel, K. Seiler: "Zusatz extrudierter Weizenmehle bei der Herstellung von Stuten und Stullen," Getreide Mehl Brot **40** (1986)
13. W. Seibel, K. Seiler: "Wirkung von extrudierten Weizenmehlen auf die Qualitätseigenschaften von feinen Backwaren," *Getreide Mehl Brot* **40** (1986)
14. R. Fast, B. Caldwell: *Breakfast Cereals and how they are made*, Association of Cereal Chemists, St. Paul, MN, 2000.
15. B. O. Juliano: Rice in Human Nutrition, FAO/IRRI, FAO Food and Nutrition Series No. 26, FAO, Rome 1993.
16. R. Bierlen, E. Wailes, E. Cramer: *The MERCOSUR Rice Economy*, Arkansas Experiment Station Bulletin. No. 954, University of Arkansas, USA, 1997.
17. UNICEF: State of World's Children, New York 1991.
18. M. C. Kik: *Effect of Milling, Processing, Washing, Cooking and Storage on Thiamine, Riboflavin and Niacin in Rice*, Agricultural Experimentation Station, Arkansas, USA, 1945.
19. P. Pillaiyar: "Rice Parboiling Research in India," *Cereal Foods World* **35** (1990) 225–257.

Ionic Liquids

G. Wytze Meindersma, Process Systems Engineering Group/Department of Chemical Engineering and Chemistry, Eindhoven University of Technology, Eindhoven, The Netherlands

Matthias Maase, BASF Aktiengesellschaft/Global New Business Development/Chemical Intermediates for Industrial Applications, Ludwigshafen, Germany (Chap. 4.1.1.2)

André B. De Haan, Process Systems Engineering Group/Department of Chemical Engineering and Chemistry, Eindhoven University of Technology, Eindhoven, The Netherlands

1. Introduction 2
2. Physical and Chemical Properties 3
2.1. Introduction 3
2.2. Melting Points 3
2.3. Density 4
2.4. Viscosity 4
2.5. Thermal stability 4
2.6. Electrochemical Window 4
3. Synthesis/Production 5
3.1. Lewis Acid-Based ILs 6
3.2. Anion Metathesis 6
3.2.1. Synthesis of Chloride-Free ILs . . . 7
3.2.2. Preparation with Microwaves 7
3.3. Producers 7
4. Applications 7
4.1. Potential applications 8
4.1.1. Solvents for Synthesis and Catalysis 8
4.1.1.1. Applications in Reactions and Catalysis 8
4.1.1.2. Solvents for Catalysis 8
4.1.1.3. Polymer Synthesis 9
4.1.1.4. Biocatalysis 9
4.1.2. Electrochemical Applications . . . 9
4.1.3. Analytical Applications 9
4.1.4. Separations 9
4.1.4.1. Liquid Separations 9
4.1.4.2. Gas Separations 10
4.1.5. Fluid Applications: Thermal Fluids and Lubricants 10
4.2. Applications in Pilot Plants and Industry 10
4.2.1. Reactions 10
4.2.1.1. Acid Scavenging 10
4.2.1.2. Production of 2,5-Dihydrofuran . . 12
4.2.1.3. Chlorination of Alcohols/Cleavage of Ethers 12
4.2.1.4. Hydrosilylation 12
4.2.1.5. Fluorination 13
4.2.1.6. Dimerization and Oligomerization of Olefins 13
4.2.2. Separations 13
4.2.2.1. Extractive Distillation 13
4.2.2.2. Extraction of Aromatic Hydrocarbons 14
4.2.3. Electroplating/Polishing 15
4.2.4. ILs as Performance Chemicals . . . 15
4.2.4.1. Compatibilizers for Pigment Pastes 15
4.2.4.2. Antistatic Additives for Cleaning Fluids 15
4.2.5. ILs in Gas Processing 16
4.2.5.1. Storage of Gases 16
4.2.5.2. Ionic Compressor 16
5. Analysis 16
6. Toxicology and Occupational Health 17
6.1. (Eco)toxicology 17
6.2. Mutagenicity 18
6.3. Safety and Corrosion 18
7. References 18

The potential to use ionic liquids as novel solvents or fluids for a diverse range of applications has become increasingly apparent. The intrinsic nonvolatile nature of ILs provides an opportunity to reduce, or even completely eliminate, hazardous and toxic emissions to the atmosphere, thus providing the promise for significant environmental benefits. In synthesis and catalysis, ILs have been used as solvents (or solvents and catalysts), with the greatest current effort on using the ILs as alternatives to volatile organic compounds (VOCs). Electrochemical studies have utilized the fact that ILs are liquids rather than solids to provide liquid

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.l14_l01

electrolytes without needing to add an additional solvent. Ionic liquids are also used in separations, replacing volatile solvents. The number of applications on pilot-plant and commercial scale is still limited, but growing.Although ionic liquids are also known as "green solvents", this is not always true. They can be corrosive, flammable, or toxic. Due to their nonvolatile nature, ionic liquids are generally considered as having a low impact on the environment and human health, and thus recognized as solvents for green chemistry. However, the impact of ionic liquids on aquatic ecosystems is important given their mild to high solubility in water. Before ionic liquids will be widely used in industry, the effects of ILs on the aquatic environment must be known.

Nomenclature of ionic liquids

Cations:
[Rmim] 1-R-3-methylimidazolium
[Rbupy] 4-R-*N*-butylpyridinium
[mRpyrr] *N*-methyl-*N*-R-pyrrolidinium
[N4111] butyl-trimethyl ammonium
[P4111] butyl-trimethyl phosphonium
[S4111] butyl-trimethyl sulfonium
R: H hydrogen
m methyl
e ethyl
p propyl
b butyl
h(e) hexyl
o(c) octyl
Anions
[Tf_2N] bis(trifluoromethylsulfonyl)imide $[(CF_3SO_2)_2N]^-$
[triflate] trifluoromethanesulfonate $[CF_3SO_3]^-$
[tosylate] toluene sulfonate $[CH_3C_6H_4SO_3]^-$
[trifluoroacetate] $[CF_3CO_2]^-$

1. Introduction [1 – 9]

The common definition of ionic liquids (ILs), or room temperature ionic liquids (RTILs), is that they are liquids that are composed entirely of ions and are fluid below 100 °C. Heating normal salts, such as sodium chloride (NaCl, *mp* 801 °C), to high temperature produces also a liquid, that consists entirely of ions, but this is a molten salt and not defined as an ionic liquid. Ionic liquids are organic salts that are liquid at low temperatures (< 100 °C) and consist of large bulky and asymmetric organic cations based on methylimidazolium [Rmim], *N*-butylpyridinium [R−N−bupy], *N*-methyl−*N*-alkylpyrrolidinium, quaternary ammonium, quaternary phosphonium ions (Fig. 1.), and many others; and anions such as hexafluorophosphate $[PF_6]^-$, tetrafluoroborate $[BF_4]^-$, alkylsulfates $[RSO_4]^-$, alkylsulfonates $[RSO_3]^-$, halides as chloride $[Cl]^-$, bromide $[Br]^-$ or iodide $[I]^-$, nitrate $[NO_3]^-$, sulfate $[SO_4]^-$, aluminum chloride $[AlCl_4]^-$, triflate $[CF_3SO_3]^-$, bis(trifluoromethylsulfonyl)imide $[(CF_3SO_2)_2N = Tf_2N]^-$, etc. [10 – 19].

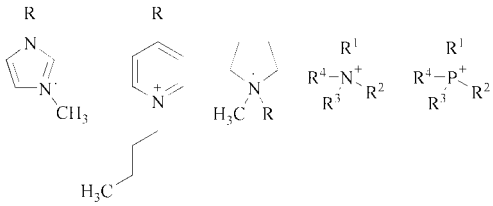

Figure 1. Structure of 1-R-3-methylimidazolium, 4-R-*N*-butylpyridinium, *N*-methyl-*N*-R-pyrrolidinium, quaternary ammonium and quaternary phosphonium cations

The R group of the cation is variable, typical are alkyl chains such as methyl, ethyl, butyl, etc., but can also be any of a variety of other functional groups (e.g., fluoroalkyl, alkenyl, methoxy). The variability of the anion and R groups in the imidazolium, pyridinium, pyrrolidinium, ammonium or phosphonium cations may be used to adjust the properties of the ionic liquids. Therefore, the possibility arises to optimize the ionic liquid for a specific application by stepwise tuning the relevant solvent properties. For this reason ionic liquids have been referred to as "designer solvents" in several publications [13]. Presently, there are more than 1000 different ionic liquids available, but it has been demonstrated [20] that there will be at least 10^6 simple ionic liquids, that can be easily prepared in the laboratory. There will be 10^{12} binary combinations of these and 10^{18} ternary systems possible. It is entirely plausible that to generate the required combination of reactivity, solubility and viscosity, industrial applications will need to work with ternary systems [21].

Ionic liquids have a wide liquidus range, ranging from 75 °C to 300 °C, which allows a

better kinetic control on reactions. The liquidus region is defined as the range of temperatures between the normal freezing point and boiling point. Thus, water has a 100 °C liquidus range. The narrow liquidus range of common solvents means that they are all relatively volatile at process conditions.

Due to the fact that an ionic liquid is a salt, it has a negligible vapor pressure. Therefore, ionic liquids are not volatile at process conditions. Furthermore, they are nonflammable, have a high thermal stability and a high ionic conductivity. These properties permit their use in many fields. They are used as reaction media to substitute VOCs, as homogeneous catalysts, and as extractive media in liquid – liquid extraction processes [1 – 19]. However, the use of ionic liquids also has some disadvantages: the physical properties are not always known, their viscosity is usually higher than that of common solvents and their toxicity is unknown. For a successful application of ionic liquids in industrial processes, these aspects must be taken into consideration.

2. Physical and Chemical Properties

2.1. Introduction

The notable characteristics of ionic liquids are their nonmeasurable vapor pressure (however, several authors have actually shown in a recent paper that there is a small but non negligible vapor pressure [22]), nonflammability (decomposition products of ILs can, however, be flammable [23]), thermal stability, wide liquid range, and solvating properties for diverse kinds of materials. Another important feature of ionic liquids is their designability: miscibility with water or organic solvents can be tuned through side chain lengths on the cation and choice of anion. Furthermore, they can be functionalized to act as acids, bases, or ligands.

The change of anion dramatically affects the chemical behavior and stability of the ionic liquid, and the change of cation has a profound effect on the physical properties, such as melting point, viscosity, and density [24 – 28]. Cations are typically big, bulky and asymmetric, accounting for the low melting points. Increasing the length of an alkyl chain tends to decrease water solubility by increasing the hydrophobicity of the cation. In addition, chemical and physical properties can be changed dramatically by the choice of anion. For example, imidazolium salts with halide, acetate, nitrate and trifluoroacetate anions are totally miscible with water, while $[PF_6]^-$ and $[(CF_3SO_2)_2N]^-$ imidazolium salts are immiscible, and $[BF_4]^-$ and $[CF_3SO_3]^-$ imidazolium salts can be totally miscible or immiscible depending on the substituents on the cation [26]. The anions are chosen to control the chemistry [29] and the cations are chosen to determine the physical properties [27]. In this way, a vast range of ionic liquids can be created. The number of possible anions [30 – 33] and cations [34 – 36] is rapidly increasing.

Ionic liquids possess the following desirable properties [37]:

- Liquidus range of 300 °C (−96 – +200 °C)
- Excellent solvents for organic, inorganic, and polymeric materials
- Acidic compositions are superacids ($pK_a \approx -20$)
- Some are water-sensitive; others are hydrophobic and air-stable
- Thermally stable under conditions up to 500 °C
- Easy to buy and simple to prepare
- No measurable vapor pressure
- Nonflammable
- Exhibit Brønsted, Lewis, Franklin, and 'super' acidity
- Highly solvating — therefore low volumes used, implying process intensification
- Catalysts as well as solvents
- Highly selective reactions

2.2. Melting Points [9]

Ionic liquids have been defined to have melting points below 100 °C and most of them are liquid at room temperature. Both cations and anions contribute to the low meting points of ionic liquids. The increase in anion size leads to a decrease in melting point [16]. For example, the melting points of 1-ethyl-3-methylimidazolium type ionic liquids with different anions, such as $[BF_4]^-$, $[Tf_2N]^-$ and $[C_2H_5SO_4]^-$ are 15 °C [38], −3 °C [32] and −20 °C, respectively (see Table 1). Size and symmetry of the cations make an important impact on the melting points

of ionic liquids and melting point generally increases with increasing substituents chain length. Large cations and increased asymmetric substitution results in a melting point reduction [39]. This can be readily seen by examining the phase diagrams for the hexafluorophosphate salts [24, 25], where the melting point shows a pronounced dependence on the chain length.

2.3. Density [9]

Ionic liquids in general are denser than water with values ranging from 1 to 1.6 g/cm^3 and their densities decrease with increase in the length of the alkyl chain in the cation [41], as is seen in Table 1, where densities of some ionic liquids and common solvents are shown. The order of increasing density for ionic liquids composed of a single cation is: $[CH_3SO_3]^- \approx [BF_4]^- < [CF_3CO_2]^- < [CF_3SO_3]^- < [C_3F_7CO_2]^- < [(CF_3SO_2)_2N]^-$ [51]. The densities of ionic liquids decrease almost linearly with increasing temperatures [46].

The densities of ionic liquids are also affected by the identity of anions. For example, the densities of 1-butyl-3-methylimidazolium type ionic liquids with different anions, such as $[BF_4]^-$, $[PF_6]^-$, and $[Tf_2N]^-$ are 1.2 g/cm^3, 1.37 g/cm^3 and 1.43 g/cm^3, respectively (Table 1).

2.4. Viscosity [9]

Generally, ionic liquids are more viscous than common molecular solvents and their viscosities are ranging from 10 mPa · s to about 500 mPa · s at room temperature. The viscosities of some popular air- and water-stable ionic liquids and common solvents at room temperature are summarized in Table 1. The viscosity of ionic liquids is determined by van der Waals forces and hydrogen bonding. Alkyl chain lengthening in the cation leads to an increase in viscosity [32], due to stronger van der Waals forces between cations leading to increase in the energy required for molecular motion. Also, the ability of anions to form hydrogen bonding has a pronounced effect on viscosity. The fluorinated anions such as $[BF_4]^-$ and $[PF_6]^-$ form viscous ionic liquids due to the formation of complexes with hydrogen bonding donors [40]. Electrostatic forces may also play an important role. In general, all ionic liquids show a significant decrease in viscosity as the temperature increases [26, 27, 46, 47, 52]. Ionic liquids should be free of impurities or moisture, because contamination of the ILs with chloride led to an increase of the viscosity [26, 27], see for instance the viscosities of several ILs in Table 1, measured by different authors.

2.5. Thermal stability

Ionic liquids can be thermally stable up to 500 °C. The thermal stability of ionic liquids is limited by the strength of their heteroatom – carbon and their heteroatom – hydrogen bonds, respectively [16]. The ionic liquids 1-ethyl-3-methylimidazolium tetrafluoroborate, 1-butyl-3-methylimdazolium tetrafluoroborate and 1,2-dimethyl-3-propylimidazolium bis(trifluorosulfonyl)imide are stable up to temperatures of 445, 423 and 457 °C, respectively [53]. However, the long-term thermal stability of ionic liquids containing dicyanamide or tricyanomethide anions is significantly lower than might be expected from temperature-ramped thermogravimetric analysis (TGA) studies, and these anions, combined with N-based cations, yield polymeric products during thermal decomposition [54].

2.6. Electrochemical Window [9]

The electrochemical window is an important property and plays a key role in using ionic liquids in electrodeposition of metals and semiconductors. By definition, the electrochemical window is the electrochemical potential range over which the electrolyte is neither reduced nor oxidized at an electrode. This value determines the electrochemical stability of solvents. As known, the electrodeposition of elements and compounds in water is limited by its low electrochemical window of only about 1.2 V. On the contrary, ionic liquids have significantly larger electrochemical windows, e.g., 4.15 V for [bmim][PF_6] at a platinum electrode [55], 4.10 V for [bmim][BF_4] [55] and 5.5 V for [bmpyrr][Tf_2N] at a glassy carbon electrode [56]. In general, the wide electrochemical windows of ionic liquids have opened the door

Table 1. Melting points, densities and viscosities of selected liquids at 20 °C

Liquid	*mp*, °C	Density, g/mL	Viscosity, mPa · s
Ethanol	−117.3	0.7893	1.200
Diethyl ether	−166.2	0.7137	0.222
Toluene	−95	0.8669	0.590
Hexane	−95	0.666	0.326
Water	0	0.9982	1.002
Dimethyl sulfoxide	18.4	1.1014	2.14
Ethylene glycol	−11.5	1.088	19.9
Glycerol	20	1.261	1490
[bmim]PF_6	11.0 [42]	1.3705 [43]	375.9 [43]
	−8.0 [44]	1.36 [44]	312.0 [44]
	10.0 [45]	1.3727 [27]	371.0 [27, 45]
		1.368 [46] (25 °C)	201.0 [47]
[bmim]BF_4	−81.0 [45]	1.2048 [42]	109.2 [43]
	−82.0 [44]	1.2077 [27]	112.0 [47]
			154.0 [26]
		1.17 [44]	233.0 [44]
[emim]BF_4	12 – 12.5 [38]	1.24 [38, 49]	37.7 [38, 49]
	15.0 [48, 49]	1.2479 [27]	66.5 [26, 45]
[bmim]Tf_2N	−2.0 [42]	1.4431 [43, 44]	59.8 [43]
	−4.0 [32, 44, 49]	1.429 [32, 49]	52.0 [32, 44, 49]
	−6.0 [46]	1.4386 [42]	58.0 [47]
		1.436 [46] (25 °C)	
[emim]Tf_2N	−17.0 [42]	1.5235 [43]	40.1 [43]
	−3.0 [32, 44, 49]	1.52 [32, 44, 49]	34.0 [32, 44, 49]
		1.5213 [42]	
		1.519 [46] (25 °C)	
[emim]$C_2H_5SO_4$	−20	1.2430 [43]	107.7 [43]
[N4111]Tf_2N*		1.3966 [43]	140.7 [43]
[mebupy]BF_4	< -50	1.18	291.3
[mebupy]BF_4 with 327 ppm water			246. [50]

* N4111 = butyltrimethylammonium

to electrodeposit metals and semiconductors at room temperature which were formerly obtained only from high temperature molten salts. For example, Al, Mg, Si, Ge, and rare earth elements can be won from room temperature ionic liquids. The thermal stability of ionic liquids allows to electrodeposit Ta, Nb, V, Se, and presumably many other elements at elevated temperature.

3. Synthesis/Production [14, 16, 57]

The first ionic liquid, ethylammonium nitrate, was produced in 1914 [58]. This compound was formed by the addition of concentrated nitric acid to ethylamine, after which water was removed by distillation to give the pure salt, which was liquid at room temperature (*mp* 12 °C). The protonation of suitable starting materials, such as amines and phosphines, still represents the simplest method for the formation of ionic liquids, but, unfortunately, it can only be applied for a few useful salts. Most ILs are formed from cations that do not contain acidic protons.

The synthesis of ILs can generally be split into two parts: formation of the desired cation, and anion exchange to form the desired product (see Fig. 2.) [16]. Quaternization of an amine or phosphine, for example, is the initial step in the synthesis of the cation of an ionic liquid [59 – 61]. If it is not possible to form the desired anion directly by the quaternization reaction, a further step is required (synthesis steps IIa or IIb in Fig. 2.).

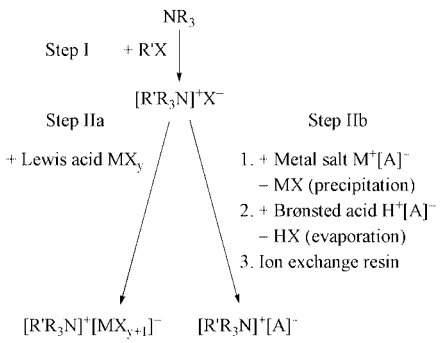

Figure 2. Typical synthesis path for the formation of ionic liquids [0]

The anion exchange of ionic liquids can be divided into three distinct paths: direct treatment of halide salts, such as an ammonium halide $[R'R_3N]^+X^-$, with Lewis acids, MX_y, most notably $AlCl_3$ (synthesis step IIa, Fig. 2.) [59, 62 – 74], the formation of ionic liquids by metathesis (synthesis step IIb, Fig. 2.), and acid – base neutralization reactions.

3.1. Lewis Acid-Based ILs

In the first case (step IIa), several anion species are often present in equilibrium, which depends on the ratio of the two components $[R'R_3N]^+X^-$ and MX_y (Eq. **??**).

The direct treatment of 1-ethyl-3-methylimidazolium chloride ([emim][Cl]) with different ratios of Lewis acids, mostly $AlCl_3$, can form the following ionic liquids: [emim][$AlCl_4$], [emim][Al_2Cl_7] or [emim][Al_3Cl_{10}]. Other Lewis acids include $AlEtCl_2$ [65, 66], BCl_3 [67] CuCl [68] and $SnCl_2$ [69]. These kind of ionic liquids form the first generation ionic liquids. A major disadvantage of these ILs is that they are water-sensitive and can decompose in the presence of water, forming HCl.

3.2. Anion Metathesis

The second route, metathesis, can be carried out by the addition of a metal salt $M^+[A]^-$, for instance a Group 1 metal, silver, or ammonium salt (with precipitation of M^+X^-), over an ion exchanger [70] or by displacement of the halide ion by a strong acid $H^+[A]^-$ (with the release of H^+X^-) (synthesis step IIb, Fig. 2.). Many alkylammonium halides are commercially available or they can be prepared simply by the reaction of the appropriate halogenoalkane and amine. Preparation of the pyridinium and imidazolium halides can be achieved similarly [75, 76]. The synthesis of [emim][Cl] was reported in1982 [59] and the synthesis of [bmim][Cl] in 1997 [77].

In 1992, the first of the new ionic liquids, [emim][BF_4], was prepared via metathesis of [emim][I] with $AgBF_4$ in methanol [48]. This salt has a melting point of 12 °C and may be prepared more cheaply using [NH_4][BF_4] in acetone [38]. This ease of preparation, together with its relative moisture stability and its immiscibility with a number of organic solvents is leading to its increasing use in biphasic catalysis. The preparation of [emim][PF_6] was accomplished by reaction of [emim][Cl] with HPF_6 [78, 79]. This salt has a melting point of 60 °C, which makes it slightly less attractive than the $[BF_4]^-$ salt. Examples of ILs prepared by anion metathesis are ILs with $[PF_6]^-$ [12, 24, 78, 79]; $[BF_4]^-$ [24, 48, 78 – 80]; $[CF_3SO_2)_2N]^-$ [32, 79]; $[CF_3SO_3]^-$ [32, 81]; $[CH_3CO_2]^-$ [48], $[CF_3CO_2]^-$ [48]; $[CF_3(CF_2)_3CO_2]^-$ [32]; $[NO_3]^-$ [26, 32, 79]; $[CB_{11}H_{12}]^-$ [82]; $[N(CN)_2]^-$ [83] and $[AuCl_4]^-$ [84].

Low-melting salts have been prepared by anion-exchange reactions for many years. The preparation of tetrahexylammonium benzoate, a liquid at 25 °C, was already reported in 1967 [85]. Ionic liquids based on cations such as tetrabutylphosphonium [86] and trimethylsulfonium [87] also have been prepared by similar synthetic methods.

The ILs [Rmim][PF_6] are by far the most popular ionic liquids used, because they are cheap, easy to prepare and to purify, are excellent archetypes for structural studies and are effectively free of IPR (Intellectual Property Rights) issues, but they are water-sensitive and will form HF at higher temperatures [88]. For this reason, these ILs will never be used in industry. The reaction of the $[PF_6]^-$ anion proceeds as follows:

ILs with thiocyanate, nonafluorobutanesulfonate, bis((trifluoromethyl)sulfonyl)imide, tris((trifluoromethyl)sulfonyl)methide, trifluoroacetate, and heptafluorobutanoate salts have all been prepared by metathesis reactions [32, 89, 90]. The metathesis reaction can leave the ionic liquids contaminated with a small amount of halide ions that may react with solute materials [91].

Tetraalkylammonium tetraalkylborides are usually prepared by the metathesis reaction of tetraalkylammonium bromide and lithium tetraalkylboride [92, 93]. Since both the ammonium cation and the boride anion are asymmetrically substituted, the starting materials are usually made rather than purchased. Reaction of a trialkylborane with the appropriate alkyllithium reagent in dry hexane yields the required lithium tetraalkylboride [94].

Other ionic liquids are made by the quarternization of the appropriate amine. This has been used to prepare a number of 1-Alkyl-3-ethylimidazolium trifluoromethanesulfonate salts [32]. Methyl triflate is reacted with a stoichiometric amount of the 1-alkylimidazole in 1,1,1-trichloromethane. Since the methyl triflate is sensitive to moisture, the reaction must be carried out under anhydrous conditions.

Monoalkylammonium nitrate salts are best prepared by the neutralization of aqueous solutions of the amine with nitric acid [90]. The ionic liquids are isolated by removing excess water in vacuum. In a similar reaction, tetraalkylammonium sulfonates have been prepared by mixing equimolar amounts of the sulfonic acid and the tetraalkylammonium hydroxide [95]. Excess water was also here removed in vacuum. To ensure the purity of the ionic liquids, they were dissolved in either acetonitrile or tetrahydrofuran and treated with activated charcoal for at least 24 h, and finally the organic solvent was removed in vacuum. The metathesis reaction can also be carried out entirely in organic solvents, such as CH_2Cl_2 [79] or acetone [81].

The final method for the synthesis of ionic liquids is direct combination of a halide salt with a metal halide. This is how the halogenoaluminate(III) and the chlorocuprate(I) ionic liquids are prepared [92, 96]. The chlorocuprate(I) ionic liquids are particularly sensitive to oxygen and have not found widespread use in synthesis [97].

3.2.1. Synthesis of Chloride-Free ILs

New low-cost ionic liquids containing methyl and ethyl sulfate anions can be easily and efficiently prepared under ambient conditions by the reaction of 1-alkylimidazoles with dimethyl sulfate and diethyl sulfate. The preparation and characterization of a series of 1,3-dialkylimidazolium alkyl sulfate and 1,2,3-trialkylimidazolium alkyl sulfate salts are reported in [98]. 1,3-Dialkylimidazolium salts containing at least one nonmethyl *N*-alkyl substituent are liquids at, or below, room temperature. The salts are stable, water-soluble, inherently 'chloride-free', display an electrochemical window of greater than 4 V, and can be used as alternatives to the corresponding halide salts in metathesis reactions to prepare other ionic liquids including 1-butyl-3-methylimidazolium hexafluorophosphate.

Not all ionic liquids are based around aromatic nitrogen-containing cations, there are also those with tetraalkylphosphonium cations [69, 99 – 109] and tetraalkylammonium cations [21, 110].

3.2.2. Preparation with Microwaves

ILs exhibit excellent heat-transfer properties and efficient absorption/transfer of microwaves and these properties have led to expeditions into "solvent-free" microwave-assisted synthesis routes of ILs. Moreover, the use of microwave synthesis has considerably speeded up ionic liquid synthesis, as well as making their synthesis a model of green chemistry, being stoichiometric, requiring no or low molar excess, and quantitative [111 – 116].

3.3. Producers

Since large-scale applications of ILs are essentially nonexisting, most ILs are prepared on a small scale for laboratory use only. Therefore, it is not surprising that currently most of the IL manufacturers are known as suppliers of laboratory chemicals. Producers of ILs on a ton scale are BASF, Degussa, and Solvent Innovation.

Present major producers/suppliers of ionic liquids are: Acros Organics, Belgium; BASF (imidazolium-based ILs), Germany [117]; Covalent Associates, USA [118]; C-Tri, South Korea; Cytec (phosphonium-based ILs), Canada/USA [105 – 107, 119]; Degussa, Germany; IoLiTec, Germany; Merck (EMD in the USA), Germany [120 – 122]; Ozark, USA, Sachem (ammonium based ILs), USA [123, 124]; Scionix (choline-based ILs), United Kingdom [125]; Sigma-Aldrich/Fluka; Solvent Innovation, Germany [126 – 128]; Strem Chemicals (phosphonium ILs), USA; TCI, Japan.

4. Applications

Until 1998, the number of entries with the terms "ionic liquid" or "ionic liquids" in the Chemical Abstracts was below or around twenty per year,

but this number is increasing from 45 per year in 1999 to 1255 per year in 2005 and 1717 in 2006. The total number of entries is now over 5300. Since the interest in ionic liquids has only increased after 1999, the development of ionic liquids is still mostly in the research phase. Therefore, there are only a few industrial applications known. However, there is a large field of potentially interesting applications. This section is divided in two parts: one dealing with applications still in the research phase and one with applications in the pilot-plant and industrial phase.

4.1. Potential applications

Most of the potential applications of ionic liquids are as solvents or catalysts in many kinds of chemical reactions such as Diels–Alder and Friedel–Crafts reactions, and also as very suitable solvents for biocatalysis. Applications in other fields, such as in separations, are lower in numbers.

4.1.1. Solvents for Synthesis and Catalysis [1 – 6, 14 – 17, 129 – 141]

Room-temperature ionic liquids have the potential to provide environmentally friendly solvents for the chemical and pharmaceutical industries because of their nonvolatile nature. Many reactions have been demonstrated in IL solvents, although in some cases without any obvious benefits over other solvents. The nature of the IL components can significantly influence reaction outcomes, either as catalysts, or modifying the reactivity of components. The activity of nucleophiles in ILs can be moderated by the anion of the IL [134, 135]. This 'buffering' action is effectively utilized [29], demonstrating how different reaction outcomes are obtained from IL+HNO_3 systems depending on the different anions present in the IL solvent.

4.1.1.1. Applications in Reactions and Catalysis [37]:

- Friedel–Crafts Alkylation [142 – 152]
- Heck and Suzuki coupling[153 – 156]
- Oxidation (with air and/or pure oxygen) [157]
- Cracking [158]
- Reduction
- Hydrogenation [159 – 161]
- Chiral hydrogenation
- Oligomerization
- Polymerization [162 – 165]
- Sulfonation
- Nitration
- Halogenation [166]
- Diazotization
- Diels–Alder reactions [167]
- *N*-alkylation and *O*-alkylation
- Aldol condensation
- Biocatalysis [136, 168 – 180]

4.1.1.2. Solvents for Catalysis [181]

Molten tetraalkylphosphonium and ammonium salts were investigated as reaction solvents/catalysts for industrial hydroformylation and hydrogenation reactions [69, 99 – 101]. Since ILs are stable, nonvolatile solvents and since the product may be separated by decantation or simple distillation [69], these properties are now considered fundamental to the application of ILs as solvents for clean syntheses. Many homogeneous catalysis reactions have subsequently been demonstrated [3, 4, 14 – 17, 133 – 140]. In cases where reactions proceed through ionic transition states, ILs can be used to tailor the transition metal complexes central to the catalyst systems, offer novel routes to new compounds, and opportunities for reaction control and separation of the homogeneous catalysts from the reactants, products and the substrate [182 – 184].

Palladium-catalyzed Heck couplings have been extensively investigated [153 – 156] and ILs are presumably the best solvents for these reactions [186]. Acid-catalyzed reactions [187] including Friedel – Crafts alkylations [142 – 152] and olefin alkylation, oligomerization, and polymerization have been studied [188]. Synthetic polyalphaolefins base oils and lubricants [189] can be prepared efficiently utilizing the ability to control acidity of the IL catalyst and the relative solubility of feedstocks and products. The Institut Français du Petrole (IFP) has developed an IL-liquid biphasic analogue of their existing mersol X process [190] for propene and butene

dimerization. The IL system uses an IL as solvent for the Ni(II) catalyst [191] with a large decrease in catalyst consumption, increased selectivity to octene, and a significant reduction in the need for downstream neutralization of entrained catalyst in the product stream.

4.1.1.3. Polymer Synthesis

The use of ILs in polymer synthesis is a recent development and it is still limited [162 – 165]. In the last few years, however, it has been shown that ionic liquids can be advantageously used also for polymer synthesis. In radical polymerization, an increase of k_p/k_t ratio for polymerizations conducted in ionic liquids has been observed. Application of ionic liquids as solvents for atom transfer radical polymerization facilitates separation of the polymer from residual catalyst and reduces the extent of side reactions. Significant increases in the rate of polymer propagation to termination steps in radical polymerization have been shown and potential applications in the preparation of specialist polymers are suggested, especially controlling formation of block copolymers, and IL-polymer gel electrolytes.

4.1.1.4. Biocatalysis

Attention has recently been focused on the use of enzymes in ionic liquids [136, 168 – 178]. It was already noted that enzymes can be relatively stable in a mixture of ILs and water. The enzyme displayed excellent stability when suspended in the ionic liquid and the activity was equal to that observed in ethyl acetate – water.

Several groups have reported on lipase-catalyzed enantioselective transesterification of chiral alcohols in ionic liquids [178 – 180]. The kinetic resolution of 1-phenylethanol with nine different lipases in ten different ionic liquids was investigated [178]. Good activities and, in many cases, improved enantioselectivities were observed compared with the same reaction in methyl *tert*-butyl ether (MTBE). Based on these initial studies the use of enzymes in ionic liquids would appear to open up a new field of nonaqueous enzymology. Ionic liquids could have added benefits for performing biotransformations with highly polar substrates, e.g., carbohydrates and amino acids, which are sparingly soluble in common organic solvents.

4.1.2. Electrochemical Applications [181]

ILs can be designed with wide electrochemical windows and high electrically conductivity [192]. When combined with wide liquid ranges and thermal stability, ILs can be used in place of traditional solvent-based electrolytes for a wide range of metal-purification, plating, and refining processes, increasing the scope of electrochemical reactions. Many IL-based electrochemical devices have been developed, including batteries [38, 59, 193 – 196], capacitors [197 – 199], and solar cells [33, 200, 201]. IL electrolytic systems are also useful for electrosynthesis [202] and electrodeposition of metals [203 – 206].

4.1.3. Analytical Applications

ILs are increasingly used in analytical chemistry, such as in chromatography, MS, spectroscopy, and electrochemistry [90, 207 – 213]. They are, for instance, employed as stationary phases in gas – liquid chromatography, because of their high viscosity, high thermal stability, low vapor pressure and wetting ability toward common supports such as fused-silica capillary tubing.

4.1.4. Separations

4.1.4.1. Liquid Separations

The use of ionic liquids in separations is presently mostly in the extraction of metal ions with [bmim][PF_6], [hmim][PF_6] and other $[PF_6]^-$ based ILs and with $[Tf_2N]^-$ based ILs [214 – 218]; of alcohols using [Rmim][PF_6] (R = butyl, pentyl, heptyl, or octyl) and [emim][Tf_2N] [219 – 221]; separation of alcohols and alkanes or alkenes with [omim][Cl], [hmim][BF_4] or [hmim][PF_6] [222, 223]; desulfurization of oils with [emim][$AlCl_4$], [bmim][$AlCl_4$], [bmim][BF_4], [bmim][PF_6], trimethylamine hydrochloride/$AlCl_3$, [emim][ethyl sulfate] and [bmim][octyl sulfate] [224 – 227]; and separation of ethers from ethanol with [omim][Cl]

and [bmim] [trifluoromethanesulfonate] [228, 229]. It is also possible to separate compounds from each other by selective transport by using supported liquid membranes based on ionic liquids, such as [bmim][PF_6] [230 – 233]. Environmental pollutants, such as aromatic and polycyclic aromatic hydrocarbons, can be extracted from aqueous solutions with ionic liquids [bmim][PF_6] and [omim][PF_6] [233]. There are only a few publications concerning extraction of aromatic hydrocarbons from mixtures of aromatic and aliphatic hydrocarbons, notably with [emim][I_3], [bmim][I_3], [emim][$(CF_3SO_2)_2N$], [omim][Cl] and trimethylamine chloride/$AlCl_3$, [hmim][BF_4] and [hmim][PF_6] [234 – 238].

High toluene/heptane selectivities were obtained with ionic liquids [mebupy][BF_4], [mebupy][CH_3SO_4], [bmim][BF_4] at 40 °C and 75 °C and with [emim][tosylate] at 75 °C. With [mebupy][BF_4], also extraction experiments with other aromatic/aliphatic combinations were carried out and high aromatic/aliphatic selectivities were obtained [239 – 244].

Using salts to break azeotropes by interacting preferentially with one component is a well established industrial procedure. The ability of ILs to break azeotropes has been described for ethanol – water, acetone – methanol, THF – water and acetic acid – water mixtures [235, 238].

It is remarkable that most of the ionic liquids used for extraction of the compounds mentioned are PF_6^--containing imidazolium-based ionic liquids. The reason is that these ionic liquids are very versatile and easy to prepare, although HF formation is likely [88].

4.1.4.2. Gas Separations

Ionic liquids can also be used for gas separations, mainly the removal of CO_2 from several gases [245 – 253], but also for ethylene – ethane or propene – propane separations [246, 254 – 259]. The CO_2 solubility is larger in ILs with anions such as $[Tf_2N]^-$, and $[methide]^-$, which contain fluoroalkyl groups, regardless whether the cation is imidazolium, pyrrolidinium, or tetraalkylammonium. These results suggest that the nature of the anion has the most significant influence on the gas solubilities [245, 246]. An increase in the alkyl chain length on the cation increases the CO_2 solubility marginally. Oxygen is hardly soluble in several ionic liquids and the solubility of CO is below the detection limit [246]. For the separation of hydrocarbon gases it was found that the solubility increased as the number of carbons increased, and also when the number of carbon double bonds increased [254 – 256].

4.1.5. Fluid Applications: Thermal Fluids and Lubricants [181]

The possible use of ILs as *thermal heat-transfer fluids* has been investigated at a research level [53, 260]; heat capacity, thermal stability and liquidus range are potentially competitive with commercial diphenyl ether and polysiloxane heat-transfer fluids.

The similarity in appearance of most ILs to oils makes their possible use as *lubricants* an obvious area of interest. Initial studies show significant promise as lubricants for a range of metal and ceramic interfaces [261 – 264]. Researchers from Alcoa [265] have also recently reported an assessment of two ILs as base oils for rolling aluminum sheet. The ILs performed comparably to current blended polyalphaolefins although problems with the poor solubility of traditional lubricant additives were anticipated. In each case here, the range of available ILs with differing physical characteristics means that appropriate IL for the specific application can be selected.

4.2. Applications in Pilot Plants and Industry

An overview of applications of ionic liquids on pilot plant and industrial scale is given in Table 2.

4.2.1. Reactions

4.2.1.1. Acid Scavenging

Since 2002, BASF is operating the first commercial multi-ton process employing ionic liquids [266 – 272], called the BASIL (Biphasic Acid Scavenging utilizing Ionic Liquids) process. The BASIL process is a general solution for all kinds of acid scavenging problems [268, 273].

Table 2. Overview of pilot plant and industrial applications of ionic liquids

Company	Process	IL is acting as . . .	Scale
BASF	acid scavenging	auxiliary	commercial
Eastman Chemical	production of 2,5-dihydrofuran	catalyst	commercial until 2004
BASF	chlorination	solvent	commercial
Eli Lilly	cleavage of ethers	catalyst/reagent	pilot
Degussa	hydrosilylation	solvent	pilot
Arkema	fluorination	solvent	pilot
IFP	olefin dimerization	solvent	pilot
Chevron Phillips	olefin oligomerization	catalyst	pilot
BASF	extractive distillation	extractant	pilot
University of Twente	extraction	extractant	pilot
Scionix	electroplating (Cr)	electrolyte	pilot
Degussa	compatibilizer	performance additive	commercial
Iolitec/Wandres	cleaning fluid	performance additive	commercial
Air Products	storage of gases	liquid support	pilot
Linde	gas compression	liquid piston	pilot

The BASIL process is utilized in a plant producing diethoxyphenylphosphine, a photoinitiator intermediate prepared by the reaction of dichlorophenylphosphene with ethanol. The reaction generates HCl as a byproduct, which must be removed to avoid an unwanted side reaction (see Fig. 3.).

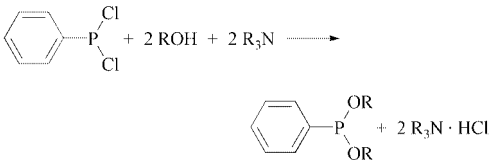

Figure 3. Synthesis of diethoxyphenylphosphine

The acid is normally removed by adding a tertiary amine, but this generates an ammonium salt slurry that is difficult to filter from the reaction mixture. BASF adds now 1-methylimidazole, which forms 1-methyimidazolium chloride (*mp* 75 °C) (Fig. 4.).

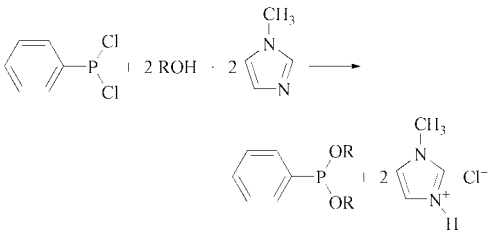

Figure 4. Synthesis of diethoxyphenylphosphine with 1-methylimidazole

After the reaction, two phases are formed, which can easily be separated: an upper phase of pure product and a lower phase of pure IL. The IL formed is deprotonated, regenerating the 1-methyimidazole for reuse. The space – time yield for the formation of an alkoxyphenylphosphine was thus increased from 8 kg $\cdot$ m^{-3} $\cdot$ h^{-1} to 690 000 kg $\cdot$ m^{-3} $\cdot$ h^{-1}, a factor of 80 000 increased productivity. The new, very small jet reactor is shown in Figure 5. BASF has successfully started a dedicated plant using this jet stream reactor technology end of 2004.

Figure 5. BASIL process in a little jet reactor (courtesy of BASF AG)

4.2.1.2. Production of 2,5-Dihydrofuran

In 1996, Eastman Chemical launched a cocatalyst ionic liquid system for the synthesis of 2,5-dihydrofuran. This compound is an intermediate in the production of many commodity, specialty, and fine chemicals. The continuous liquid-phase process uses a trialkyltin iodide Lewis acid and a tetraalkylphosphonium iodide ionic liquid to catalyze the rearrangement of 3,4-epoxy-1-butene to 2,5-dihydrofuran.

The continuous rearrangement process requires that the catalysts are essentially nonvolatile so that they do not codistill with the product. The cocatalyst system gives a high selectivity for 2,5-dihydrofuran and provides an efficient means for catalyst recovery [274 – 277].

The process was operated by Texas Eastman Division in a plant with a capacity of 1 400 t/a at Longview. Although the process was a technical success, the plant was shut down in December 2004 for business reasons, when Eastman exited from production of fine chemicals.

4.2.1.3. Chlorination of Alcohols/Cleavage of Ethers

In principle, exchange of the OH groups of alcohols by chlorine can be effected by reaction with HCl gas. Unfortunately, in the case of diols usually less than 100 % conversion is achieved and the reaction stops after formation of cyclic or open-chain ethers. Surprisingly, the reactivity profile completely changes when the reaction is performed in an ionic liquid. Obviously, either the nucleophilicity of HCl or the nucleofugicity of the leaving group (water) is drastically increased. Apparently, the ether side products are cleaved, allowing for a further reaction to the desired bischlorinated product. Figure 6. shows how selectivities improve when the chlorination of butanediol is performed in an ionic liquid rather than in an organic solvent [278]. Butanediol is completely soluble in the ionic liquid. When the reaction proceeds, a second, organic phase is formed, which consists of the reaction product 1,4-dichlorobutane. After complete conversion, the organic phase is just separated off. Water which is formed during the reaction is distilled off from the remaining ionic liquid which can be used for the next run without any further work-up.

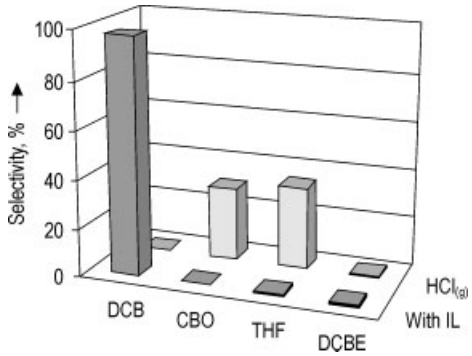

Figure 6. Higher selectivities for 1,4-dichlorobutane (DCB) are achieved if the chlorination of butanediol with HCl gas is performed in an ionic liquid rather than in the pure alcohol. Side products usually are ethers (tetrahydrofuran, THF, dichlorobutyl ether, DCBE) or the monochlorinated product 1-chlorobutane-4-ol (CBO).

In a process by Eli Lilly aromatic methoxy ethers are cleaved in a pyridinium hydrochloride melt [279]. Usually, the cleavage of aromatic methoxy ethers requires very harsh conditions and inconvenient reagents such as BBr_3 [280] or HBr in boiling acetic acid [281, 282]. 4-Hydroxyphenylbutyric acid has been produced by Eli Lilly on a 190 L pilot scale with yields of > 94 % by reacting 4-methoxyphenylbutyric acid for 2 h at 210 °C in pyridinium hydrochloride ([py][HCl]).

4.2.1.4. Hydrosilylation

Hydrosilylation is a widely used method for the synthesis of organo-modified silanes and siloxanes. The addition of Si-H to C-C double bonds is usually catalyzed by homogenous or colloidal Pt catalysts. The major drawback of this reaction is that the catalyst cannot easily be removed from the product after completion of the reaction. Degussa uses a biphasic approach to perform the hydrosilylation reaction [283, 284], where the catalyst is dispersed in the ionic liquid phase, from which the pure product separates as a new liquid phase that can be easily decanted after the reaction. The ionic liquid catalyst phase is still active and can be reused. It is important that standard hydrosilylation catalysts can be used

without further modifications. Degussa has been running this process on a pilot scale achieving conversions of > 99 %.

4.2.1.5. Fluorination

Hydrofluorocarbons (HFC) can be manufactured from chlorinated hydrocarbons by reacting them with HF to substitute fluorine for chlorine. State-of-the-art catalyst for this reaction is $SbCl_5$. However, this catalyst suffers from a reductive deactivation by formation of Sb(III) species. To overcome the loss of catalyst, chlorine is co-fed for a reoxidation of Sb(III) to Sb(V). This is technically feasible, but consumes chlorine as raw material and leads to back-chlorination of the fluorinated products and hence to significantly lower yields. Arkema has demonstrated on a pilot stage that imidazolium-based ionic liquids with an "SbF_6" anion can eliminate the deactivation problem [285]. The pilot reaction has been running for more than 1000 h exceeding selectivities of 99.5 % and without any noticeable deactivation.

4.2.1.6. Dimerization and Oligomerization of Olefins

IFP has developed an ionic liquid-based process, the biphasic Difasol technology, for the dimerization of butenes on a pilot scale [190, 191, 286 – 288]. Because of the solubility of the catalyst, based on a Ni(II) salt, in the ionic liquid phase, $[bmim]Cl/AlCl_3/C_2H_5AlCl_2$ (1:1.22:0.06 molar ratio), and the poor miscibility of the products, the production unit is essentially reduced to a continuously stirred tank reactor followed by a phase separator. With this process, a more efficient overall catalyst utilization and a significant increase in the yield of octenes is achieved. A dimer selectivity in the range of 90 to 92 % with butene conversion in the range of 80 to 85 % can be obtained with a C_4 feed containing 60 % butenes.

Chevron Phillips has found that it is possible to readily make polyalphaolefins with very high viscosity using an ionic liquid catalyst. The oligomerization reaction can thus be carried out in the absence of organic solvents which have hitherto been used as diluents for the feed [289]. The first component of the oligomerization catalyst is an aluminum halide or an alkyl aluminum halide, such as, for example, aluminum trichloride. The second component is a quaternary ammonium, quaternary phosphonium, or tertiary sulfonium ion. The process produces a synthetic lubricant base oil which is used in a variety of lubricants, including gear oils, greases, and automotive engine lubricants. The pilot process produced commercial quantities of more than 450 t of polyalphaolefins in 2004.

4.2.2. Separations

4.2.2.1. Extractive Distillation

Ionic liquids are suitable as entrainers for a whole range of azeotropic systems [238, 290, 291]. Especially if water is part of the azeotropic mixture, very high separation factors can be achieved, because ionic liquids usually are hygroscopic materials with a strong affinity to water. Figure 7. displays the classic vapor – liquid diagram for THF – water. The dots indicate the change after the addition of ionic liquid. To afford a sufficient separation, the amount of ionic liquid added has to be in the range of 30 to 50 wt %.

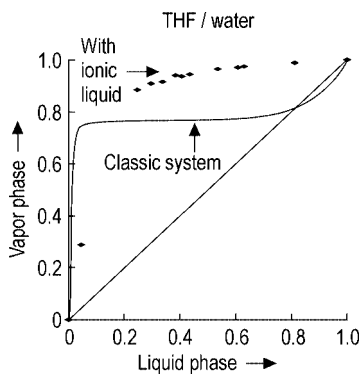

Figure 7. Equilibrium phase diagram for the system THF – water. The solid line shows the classic azeotropic mixture. The dots indicate how the azeotrope has been broken after addition of the ionic liquid. The amount of THF in the vapor phase is always larger than in the liquid phase.

As the ionic liquid has no vapor pressure, the entrainer itself need not be distilled, hence a sec-

ond separation column is not required and energy can be saved. A benchmark calculation has revealed that a saving potential of about 37 % for energy cost and 22 % for the investment can be achieved. BASF has run an extractive distillation process in a pilot plant continuously for three months. Although the ionic liquid faced a severe thermal treatment of about 250 °C in the recycling step, it fully kept its performance without a purge.

4.2.2.2. Extraction of Aromatic Hydrocarbons

Extraction of toluene from a mixture of toluene and heptane with the ionic liquid [mebupy][BF_4] was carried out in a pilot plant at the University of Twente [292 – 294]. The pilot extraction unit was a Rotating Disc Column with an inside diameter of 60 mm, 40 stirred compartments and a total active length of 1800 mm (Fig. 8.). The performances of sulfolane and [mebupy][BF_4] as

Figure 8. Pilot Rotating Disc Column for extraction of toluene, inside diameter 60 mm, 40 stirred compartments and a total active length of 1800 mm

the solvent were compared for the extractive removal of 10 wt % toluene from a *n*-heptane feed. When dispersed, the ionic liquid [mebupy][BF_4] appeared to be a better solvent for the toluene – heptane separation than sulfolane. Due to the higher toluene distribution coefficient (0.44 vs. 0.3 for sulfolane), lower solvent-to-feed ratios are required with the IL to extract the same amount of toluene, while compared to sulfolane a 10 % higher flux (volumetric throughput) and similar mass transfer efficiency was obtained. The best performance (highest toluene extraction efficiency) was obtained at the highest rotation speed used (643 rpm) during the extraction experiments. At a flux of 10 $m^3\ m^{-2}\ h^{-1}$ excellent hydrodynamic behavior was observed and three to four equilibrium stages were contained in the 1.80 m high active section of the column.

4.2.3. Electroplating/Polishing

The major disadvantage of the current process of chrome plating is that it requires the use of chromic acid-based electrolytes comprising hexavalent chromium, which is toxic and carcinogenic. Scionix, a joint venture company between the University of Leicester and Whyte Chemicals Ltd, has developed an ionic liquid containing Cr(III) salts [295], which is significantly less toxic than the carcinogenic Cr(VI) species. The process operates with > 90 % current efficiency, resulting in significantly reduced power consumption and making the ionic liquid technology a more environmentally-benign form of plating [296]. Moreover, because there are no aqueous solutions present, hydrogen evolution is negligible, hence essentially crack-free, highly corrosion-resistant deposits are possible. The process has been operated on a 50-L pilot scale and is currently being developed by a number of industrial partners.

Electropolishing of stainless steel is an effective way of increasing corrosion resistance, and decreasing wear in addition to the obvious aesthetic benefits. Scionix had developed a concept to form eutectic-based ionic liquids, from a variety of quaternary ammonium salts, such as choline chloride, and carboxylic acids. The mixture is stable because the chloride ion is complexed by the associated acid rather than mixing quaternary ammonium salts with metal halides [297, 298]. The ionic liquids allow electropolishing with high current efficiency (> 80 %), improved surface finish and improved corrosion resistance [299]. When the metal, and it is usually iron, is electrochemically oxidized, it complexes with the hydrogen-bond donor and the insoluble complex precipitates to the base of the cell. The cell liquid can be filtered periodically and the metal complex is collected. This allows for a suitable recycling of both the ionic liquid and the metal, preventing environmental emissions. The process is currently running in a 1200-L demonstrator module in Birmingham, UK.

Scionix is currently developing a commercially viable medium-to-large-scale electropolishing plant using these novel electrolytes in collaboration with UK based Anopol Ltd.

4.2.4. ILs as Performance Chemicals

4.2.4.1. Compatibilizers for Pigment Pastes

Up to now only very little attention has been paid to ionic liquids as performance chemicals. Degussa has managed to develop an application of ionic liquids as so-called compatibilizers for pigment pastes [300, 301].

Paints and coatings can be tinted by adding a small amount of a color concentrate — the pigment paste — to a white base paint. For environmental reasons, it is preferable to use water-based pigment pastes. However, the water-based systems cannot universally be used for both water- and solvent-based paints. This problem has been solved by addition of ionic liquids to the pigment paste.

4.2.4.2. Antistatic Additives for Cleaning Fluids

Wandres Micro-Cleaning GmbH has developed a cleaning system for high-value surfaces that uses moistened instead of dry filaments [302].

The liquid film (water) is brought onto the filament with the help of very small, micrometer sized water droplets, which are generated in a Venturi nozzle. To avoid electrostatic charging of the surface, a supporting salt is usually added that facilitates electrical conductivity. If sodium chloride is used as a supporting salt, a solid is

precipitated in the nozzle leading to encrustation and plugging of the system. Ionic liquids can offer a unique solution to this problem, since as liquid salts they can provide electrical conductivity without precipitation of a solid, hence without formation of deposits. The ionic liquid solution to this problem was developed and established by Iolitec in cooperation with Wandres Micro-Cleaning GmbH.

4.2.5. ILs in Gas Processing

4.2.5.1. Storage of Gases

The semiconductor manufacturing industry uses a number of hazardous specialty gases such as phosphine (PH_3), arsine (AsH_3) and boron trifluoride (BF_3) for doping, etching, and thin-film deposition. These gases are highly toxic and pyrophoric. Therefore, storage and handling is challenging and requires enormous safety efforts. For example, storage of toxic gases under high pressure in metal cylinders is often unacceptable due to the possibility of leakage or rupture of the cylinder. To overcome the risk of high-pressure storage, these gases are often stored under low pressure by adsorption to solid supports like zeolites [303]. Disadvantages of this technology are low capacities, delivery limitations and low thermal conductivity. Air Products & Chemicals have developed subatmospheric systems based on ionic liquids for storing phosphine (PH_3) and boron trifluoride (BF_3) [304, 305]. Ionic liquids have ideal properties for reversibly storing gases through chemical complexation. Due to the superior heat transfer capability of ionic liquids, gases are easier to load and to deliver, compared to solid adsorption-based systems.

To increase the affinity of the gas to the ionic liquid, a Lewis acid/Lewis base interaction is utilized. This allows for high gas storage capacities without the need for pressure. Phosphine for example is a Lewis base. Accordingly, for the absorption a Lewis acidic ionic liquid such as [bmim][Cu_2Cl_3] is used. Boron trifluoride in turn is a Lewis acid. In this case a Lewis basic ionic liquid like [bmim][BF_4] is used as liquid support (Fig. 9.).

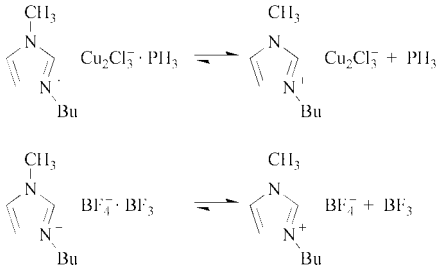

Figure 9. Adsorption/desorption of PH_3 or BF_3 with ionic liquids

Air Products has developed full scale gas packages (2.2-L cylinders), that exhibit excellent gas evolution rates, evolved gas purity, and long-term gas/liquid stability. Commercial PH_3 and BF_3 packages for manufacturing semiconductor devices have been fully qualified at several production facilities.

4.2.5.2. Ionic Compressor

Ionic liquids as so-called engineering fluids or new materials can be a functional part of devices, equipment, and machinery, that will be used in a variety of industries such as automotive, air- and spacecraft, textile, electronics, machinery, or energy. The most remarkable application in this context has been reported by Linde [306]. The so-called "ionic compressor" uses ionic liquids as "liquid piston" allowing for an almost isothermal high pressure compression of gases. Major advantages of this new technology are about ten times higher service times (i.e., operation time without maintenance) and 20 % less energy demand. Moreover, the apparatus is less complex, because it contains only eight rather than some 500 moving parts.

5. Analysis

The purity of an ionic liquid is frequently determined by NMR and other analytical techniques: ^{1}H NMR, ^{13}C NMR spectroscopy [307], mass spectroscopy, and elemental analysis are used to characterize ionic liquids. The crystal structures of IL compounds can be determined by X-ray single crystal diffraction. The powder diffraction method can be used to compare the structural similarities between the single crystal and

the microcrystalline bulk composition. Thermal properties of ionic liquids were studied using thermogravimetry/differential thermal analysis (TG/DTA) and differential thermal calorimetry (DSC) methods [308].

Solvatochromic and fluorescent probe compounds have been utilized to determine the polarity of solvents, including ionic liquids. The application of solvatochromic methods to the characterization of solvent properties of room temperature ionic liquids is discussed in [28]. Gas chromatography is a widely used method for determining thermodynamic properties of pure substances or solvent properties of binary mixtures in which one component, the sample, is generally present at infinite dilution and the other, the stationary phase, is most commonly a nonvolatile liquid. Properties of room-temperature ionic liquids and their solvation properties can thus be derived from measurements by gas chromatography [209].

The structure of ionic liquids can be determined by X-ray absorption fine structure (XAFS) [309], X-ray photoelectron spectroscopy (XPS) or low energy ion scattering (LEIS) [310].

6. Toxicology and Occupational Health

6.1. (Eco)toxicology

Although ionic liquids are also known as "green solvents", they are not always green. They can be corrosive, flammable, or toxic. Due to their nonvolatile nature, ionic liquids are generally considered as having a low impact on the environment and human health, and thus recognized as solvents for green chemistry. However, the impact of ionic liquids on aquatic ecosystems is important given their mild to high solubility in water. Before ionic liquids will be widely used in industry, the effects of ILs on the aquatic environment must be known.

Research into IL aquatic toxicity has shown them to be as toxic as or more so than many current solvents already in use [311]. Ecotoxicological tests, using the OECD standard methods on some ionic liquids, have revealed that several have a very high toxicity towards freshwater algae and the freshwater invertebrate *Daphnia magna*, about 10^4 to 10^6 times more ecotoxic than methanol for the worst cases [312]. Toxicity was greatest for ILs with eight-carbon alkyl chains attached to both imidazolium and pyridinium rings and declined with shorter alkyl chains, indicating a positive relationship between alkyl chain length and toxicity [313, 314]. ILs containing 10, 11, 12 and 14 carbon atoms in an alkoxy group show very high antimicrobial activity [315, 316]. Toxicity appeared to be related to the imidazolium cation and not to the various anions (e.g., Cl^-, Br^-, PF_6^-, and BF_4^-) [317]. Most of the commercially available anions showed no or only marginal cytotoxic effects [318]. However, for certain anions, e.g., $[(C_2F_5)_3PF_3]^-$, $[SbF_6]^-$, $[N(CF_3)_2]^-$, [bis-[1,2-benzenediolato (2-)]borate]$^-$, a significant influence on the ionic liquid cytotoxicity was demonstrated.

Besides the well-established link between toxicity and alkyl chain length on imidazolium, pyridinium, and quaternary ammonium-based ionic liquids, models predict that toxicity increases slightly with the number of nitrogen atoms in an aromatic cation ring. All other things being equal, toxicity is expected to show the trend with cation type of ammonium $<$ pyridinium $<$ imidazolium $<$ triazolium $<$ tetrazolium. In addition, toxicity is expected to decrease with ring methylation as well as with an increase in the number of negatively charged atoms on the cation. The anion plays a secondary role in toxicity for the compounds studied here, although the presence of positively charged atoms on the anion are predicted to slightly increase toxicity [319].

More data is becoming available, such as on cytotoxicity [320], toxicity to *Caenorhabditis elegans* [321], *Oocystis submarina*, and *Cyclotella meneghiniana* [322], resistance to biodegradation [323 – 325] and effects on the growth of lactic acid-producing bacteria [326, 327] and on zebra fish (*Danio rerio*) [311]. Imidazolium and pyridinium ILs inhibit the enzyme acetylcholinesterase, but bulky phosphonium ILs were less inhibitory [328].

A strategy using the structural formula of ionic liquids to generate qualitative and semi-quantitative information in regard to structure – activity relationships (SAR) is presented in [329], accompanied by (eco)toxicological rapid

screening tests [330]. Furthermore, a systematic algorithm was presented to analyze the structure of ionic liquids with respect to the molecular interaction potential within different biotic and abiotic systems [331].

6.2. Mutagenicity

Several ionic liquids were assessed for mutagenicity using the Ames test [332]. The ionic liquids tested were imidazolium based ILs ([bmim][Br], [hmim][Br] and [omim][Br]), pyridinium based ILs ([mebupy][Br], [mehepy][Br] and [meocpy][Br]), and quaternary ammonium based ILs ([4m-amm][Br], [4e-amm][Br], [4b-amm][Br] and [4h-amm][Br]). The Ames test for mutagenicity indicated that none of the ILs tested caused frameshift or missense mutations. Only some ILs, [bmim][Br] and [omim][Br], indicated trends toward potential mutagenicity at high doses, but did not meet the USEPA criteria for classification as a mutagen.

6.3. Safety and Corrosion

The nonflammability of ionic liquids is often highlighted as a safety advantage of ILs over volatile organic compounds (VOCs), but the fact that many ILs are not flammable themselves does not mean that they are safe to use near fire and/or heat sources. A large group of ILs (including commercially available ILs) are combustible due to the nature of their positive heats of formation, oxygen content, and decomposition products (see Fig. 10.) [23].

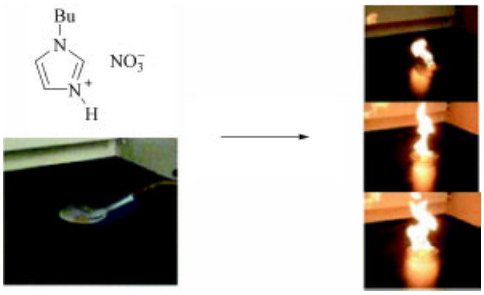

Figure 10. Ignition test of 1 g of protonated [1-bu-3-H-im][NO_3]

The corrosivity for a given metal is determined by the chemical structure of the IL cation and the nature of the anion. Negative effects were observed with tosylate and dimethylphosphate anions. Dilution with water could lead to hydrolysis of many anions used in ILs which then produce acids (e.g. HF, HCl, sulfuric or phosphoric acid) and, hence, cause acid corrosion. The best performance in all seven IL media tested was observed with stainless steel 304 [333]. Also several other ionic liquids, such as [emim]I_3 and [bmim]$_3$, are very corrosive [239], probably caused by the anion. The corrosion behavior of several metals and metal alloys (copper, nickel, AISI 1018 steel, brass, Inconel 600) exposed to a typical ionic liquid, the 1-butyl-3-methyl-imidazolium bis(trifluoromethanesulfonyl) imide, ([bmim][Tf_2N]), has been investigated by electrochemical and weight-loss methods [334]. At room temperature, there is hardly any corrosion, but at temperatures > 70 °C, the corrosion current for copper dramatically increases showing a strongly dependence on temperature.

The ILs [Rmim][PF_6] are by far the most popular ionic liquids used, because they are cheap, easy to prepare and to purify, are excellent archetypes for structural studies and are effectively free of IPR issues, but they are water sensitive and will form HF at higher temperatures [88].

7. References

General References

1. P. Wasserscheid, T. Welton (eds.): *Ionic Liquids in Synthesis*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim 2003.
2. R. D. Rogers, K. R. Seddon, S. Volkov (eds.): *Green Industrial Applications of Ionic Liquids*, NATO Science Series, II. Mathematics, Physics and Chemistry-Vol. 92, Kluwer Academic Publishers, Dordrecht 2002.
3. R. D. Rogers, K. R. Seddon (eds.): *Ionic Liquids. Industrial Applications to Green Chemistry*, ACS Symposium Series 818, American Chemical Society, Washington 2002.
4. R. D. Rogers, K. R. Seddon (eds.): *Ionic Liquids as Green Solvents. Progress and Prospects*, ACS Symposium Series 856,

American Chemical Society, Washington 2003.
5. R. D. Rogers, K. R. Seddon (eds.): *Ionic Liquids IIIA: Fundamentals, Progress, Challenges, and Opportunities. Properties and Structure*, ACS Symposium Series 901, American Chemical Society, Washington 2005.
6. R. D. Rogers, K. R. Seddon (Eds.): *Ionic Liquids IIIB: Fundamentals, Progress, Challenges, and Opportunities. Transformations and Processes*, ACS Symposium Series 902, American Chemical Society, Washington 2005.
7. ISI Essential Science Indicators, *Ionic Liquids*, http://www.esi-;topics.com/ionic-liq-uids/index.html (2004).
8. C. Chiappe, D. Pieraccini: "Review Commentary. Ionic liquids: solvent properties and organic reactivity", *J. Phys. Org. Chem.* **18** (2005), (4) 275 – 297, http://dx.doi.org/10.1002/poc.863.
9. F. Endres, S. Z. El Abedin: "Air and water stable ionic liquids in physical chemistry", *Phys. Chem. Chem. Phys.* **8** (2006) 2101 – 2116, **DOI:** 10.1039/b600519p

Specific References

10. K. R. Seddon: "Ionic Liquids for Clean Technology", *J. Chem. Tech. Biotechnol.* **68** (1997) no. 4, 351 – 356.
11. K. R. Seddon: "Ionic Liquids for Clean Technology: An Update", *Molten Salt Forum* **5-6** (1998) 53 – 62.
12. J. G. Huddleston, H. D. Willauer, R. P. Swatloski, A. E. Visser, R. D. Rogers: "Room Temperature Ionic Liquids as Novel Media for 'Clean' Liquid-liquid Extraction", *Chem. Commun.* **16** (1998) 1765 – 1766.
13. M. Freemantle: "Designer Solvents. Ionic liquids may boost clean technology development", *Chem. Eng. News* **76** (1998) no. 13, 32 – 37.
14. T. Welton: "Room-Temperature Ionic Liquids. Solvents for Synthesis and Catalysis", *Chem. Rev.* **99** (1999) no. 8, 2071 – 2083.
15. J. D. Holbrey, K. R. Seddon: "Ionic Liquids", *Clean Prod. Proc.* **1** (1999) 223 – 236.
16. P. Wasserscheid, W. Keim: "Ionic Liquids – New "Solutions" for Transition Metal Catalysis", *Angew. Chem. Int. Ed.*, **39** (2000) 3772 – 3789.
17. C. M. Gordon: "New developments in catalysis using ionic liquids", *Appl. Catal. A* **222** (2001) 101 – 117.
18. J. F. Brennecke, E. J. Maginn: "Ionic Liquids: Innovative Fluids for Chemical Processing", *AIChE Journal* **47** (2001) no. 11, 2384 – 2389.
19. J. Dupont, C. S. Consorti, J. Spencer: "Room Temperature Molten Salts: Neoteric "Green" Solvents for Chemical Reactions and Processes", *J. Braz. Chem. Soc.* **11** (2000) no. 4, 337 – 344.
20. K. R. Seddon: "Ionic liquids: Designer solvents?" in S. Boghosian, V. Dracopoulos, C. G. Kontoyannis, G. A. Voyiatzis (eds.): The International George Papatheodorou Symposium: Proceedings, (Institute of Chemical Engineering and High Temperature Chemical Processes), Patras 1999, pp. 131 – 135.
21. BP Chem Int. Ltd, WO 95 21872, 1995 (A. K. Abdul-Sada, K. R. Seddon, N. J. Stewart).
22. M. J. Earle et al.: "The distillation and volatility of ionic liquids", *Nature* **439** (2006) 831 – 834.
23. M. Smiglak et al.: "Combustible ionic liquids by design: is laboratory safety another ionic liquid myth?", *Chem. Commun.* (2006) no. 24, 2554 – 2556, **DOI:** 10.1039/b602086k.
24. C. M. Gordon, J. D. Holbrey, A. R. Kennedy, K. R. Seddon: "Ionic liquid crystals: hexafluorophosphate salts", *J. Mater. Chem.* **8** (1998) no. 12, 2627 – 2636.
25. J. D. Holbrey, K. R. Seddon: "The phase behaviour of 1-alkyl-3-methylimidazolium tetrafluoroborates; ionic liquids and ionic liquid crystals", *J. Chem. Soc., Dalton Trans.* 1999 2133 – 2140, **DOI:** 10.1039/a902818h.
26. K. R. Seddon, A. Stark, M. J. Torres: "Influence of chloride, water, and organic solvents on the physical properties of ionic liquids". *Pure Appl. Chem.* **72** (2000) no. 12, 2275 – 2287.
27. K. R. Seddon, A. Stark, M. J. Torres: "Viscosity and density of 1-alkyl-3-methylimidazolium ionic liquids", in M. Abraham, L. Moens (eds.): *Clean Solvents: Alternative Media for Chemical Reactions and Processing*, ACS Symp. Ser., Vol. 819, American Chemical Society, Washington D.C. 2002, pp. 34 – 49.
28. C. F. Poole: "Chromatographic and spectroscopic methods for the determination of solvent properties of room temperature ionic liquids", *J. Chromatogr. A* **1037** (2004) no. 1-2, 49 – 82, **DOI**: 10.1016/j.chroma.2003.10.127.
29. M. J. Earle, S. P. Katdare, K. R. Seddon, "Paradigm confirmed: The first use of ionic

liquids to dramatically influence the outcome of chemical reactions", *Org. Lett.* **6** (2004) no. 5, 707 – 710.
30. D. R. MacFarlane, S. A. Forsyth, J. Golding, G. B. Deacon: "Ionic liquids based on imidazolium, ammonium and pyrrolidinium salts of the dicyanamide anion", *Green Chem.* **4** (2002) no. 5, 444 – 448.
31. J. Golding, S. Forsyth, D. R. MacFarlane, M. Forsyth, G. B. Deacon: "Methanesulfonate and p-toluenesulfonate salts of the *N*-methyl-*N*-alkylpyrrolidinium and quaternary ammonium cations: novel low cost ionic liquids", *Green Chem.* **4** (2002) no. 3, 223 – 229.
32. P. Bonhôte, A.-P. Dias, N. Papageorgiou, K. Kalyanasundaram, M. Grätzel: "Hydrophobic, Highly Conductive Ambient-Temperature Molten Salts", *Inorg. Chem.* **35** (1996) no. 5, 1168 – 1178, **DOI**: 10.1021/ic951325x.
33. N. Papageorgiou, Y. Athanassov, M. Armand, P. Bonhôte, H. Pettersson, A. Azam, M. Grätzel: "The performance and stability of ambient-temperature molten salts for solar cell applications", *J. Electrochem. Soc.* **143** (1996) no. 10, 3099 – 3108.
34. J. Sun, D. R. MacFarlane, M. Forsyth: "A new family of ionic liquids based on the 1-alkyl-2-methylpyrrolinium cation", *Electrochim. Acta* **48** (2003) no. 12, 1707 – 1711.
35. S. Forsyth, J. Golding, D. R. MacFarlane, M. Forsyth: "*N*-methyl-*N*-alkylpyrrolidinium tetrafluoroborate salts: ionic solvents and solid electrolytes", *Electrochim. Acta* **46** (2001) no. 10-11, 1753 – 1757.
36. J. J. Golding, D. R. Macfarlane, L. Spiccia, M. Forsyth, B. W. Skelton, A. H. White: "Weak intermolecular interactions in sulfonamide salts: Structure of 1-ethyl-2-methyl-3-benzylimidazolium bis[(trifluoromethyl)sulfonyl]amide", *Chem. Commun.* 1998, no. 15, 1593 – 1594.
37. G. Fitzwater et al.: "Ionic Liquids: Sources of Innovation", Report Q002, January 2005, QUILL, Belfast, 2005, http://quill.qub.ac.uk/sources.
38. J. Fuller, R. T. Carlin, R. A. Osteryoung: "The room temperature ionic liquid 1-ethyl-3-methylimidazolium tetrafluoroborate: electrochemical couples and physical properties", *J. Electrochem. Soc.* **144** (1997) 3881 – 3886.
39. J. D. Holbrey, R. D. Rogers: "Physiochemical Properties of Ionic Liquids," in P. Wasserscheid, T. Welton (eds.): *Ionic Liquids in Synthesis*, Wiley-VCH, Weinheim 2003, pp. 41 – 55.
40. P. A. Z. Suarez, S. Einloft, J. E. L. Dullius, R. F. De Souza, J. Dupont: "Synthesis and physical-chemical properties of ionic liquids based on 1-n-butyl-3-methylimidazolium cation". *J. Chim. Phys. Phys.-Chim. Biol.* **95** (1998) 1626 – 1639.
41. K. N. Marsh, J. A. Boxall, R. Lichtenthaler: "Room temperature ionic liquids and their mixtures — a review" *Fluid Phase Equilib.* **219** (2004) no. 1, 93 – 98, **DOI**: 10.1016/j.fluid.2004.02.003.
42. C. P. Fredlake, J. M. Crosthwaite, D. G. Hert, S N. V. K. Aki, J. F. Brennecke: "Thermophysical Properties of Imidazolium-Based Ionic Liquids", *J. Chem. Eng. Data*, **49** (2004) no. 4, 954 – 964, **DOI**: 10.1021/je034261a.
43. J. Jacquemin, P. Husson, A. A. H. Padua, V. Majer: "Density and viscosity of several pure and water-saturated ionic liquids", *Green Chem.* **8** (2006) 172 – 180.
44. S. Carda-Broch, A. Berthod, D. W. Armstrong: "Solvent properties of the 1-butyl-3-methylimidazolium hexafluorophosphate ionic liquid", *Anal Bioanal. Chem.* **375** (2003) 191 – 199, **DOI**: 10.1007/s00216-002-1684-1.
45. J. G. Huddleston, A. E. Visser, W. M. Reichert, H. D. Willauer, G. A. Broker, R. D. Rogers: "Characterization and comparison of hydrophilic and hydrophobic room temperature ionic liquids incorporating the imidazolium cation", *Green Chem.* **3** (2001) 156 – 164, **DOI:** 10.1039/b103275p.
46. S. V. Dzyuba, R. A. Bartsch: "Influence of Structural Variations in 1-Alkyl(aralkyl)-3-Methylimidazolium Hexafluorophosphates and Bis(trifluoromethylsulfonyl)imides on Physical Properties of the Ionic Liquids", *ChemPhysChem* **3** (2002) no. 2, 161 – 166, **DOI**: 10.1002/1439-7641(20020215)3:2<161::AID-CPHC161 > 3.0.CO;2-3.
47. O. O. Okoturo, T. J. VanderNoot: "Temperature dependence of viscosity for room temperature ionic liquids", *J. Electroanal. Chem.* **568** (2004) 167 – 181, **DOI**: 10.1016/j.jelechem.2003.12.050.
48. J. S. Wilkes, M. J. Zaworotko: "Air and water stable 1-ethyl-3-methylimidazolium based

ionic liquids", *J. Chem. Soc., Chem. Commun.* (1992) 965 – 967, **DOI:** 10.1039/C39920000965.
49. R. Hagiwara, Y. Ito: "Room temperature ionic liquids of alkylimidazolium cations and fluoroanions", *J. Fluorine Chem.* **105** (2000) no. 2, 221 – 227.
50. J. M. Crosthwaite, M. J. Muldoon, J.-N. K. Dixon, J. L. Anderson, J. F. Brennecke: "Phase transition and decomposition temperatures, heat capacities and viscosities of pyridinium ionic liquids", *J. Chem. Thermodyn.* **37** (2005) no. 8, 559 – 568, **DOI**: 10.1016/j.jct.2005.03.013.
51. R. A. Mantz, P. C. Trulove, "Viscosity and Density of Ionic Liquids", in P. Wasserscheid, T. Welton (eds.), *Ionic Liquids in Synthesis*, Wiley-VCH, Weinheim 2003, pp. 56 – 68.
52. P. Wasserscheid, R. Van Hal, A. Bösmann: "1-*n*-Butyl-3-methylimidazolium ([bmim]) octylsulfate-an even greener ionic liquid", *Green Chem.* **4** (2002) no. 4, 400 – 404, **DOI:** 10.1039/b205425f.
53. M. E. van Valkenburg, R. L. Vaughn, M. Williams, J. S. Wilkes: "Thermochemistry of ionic liquid heat-transfer fluids", *Thermochim. Acta* **425** (2005) no. 1-2, 181 – 188, **DOI:** 10.1016/j.tca.2004.11.013.
54. T. J. Wooster, K. M. Johanson, K. J. Fraser, D. R. MacFarlane, J. L. Scott: "Thermal degradation of cyano containing ionic liquids", *Green Chem.* **8** (2006) 691 – 696, **DOI:** 10.1039/b606395k.
55. U. Shröder, J. D. Wadhawan, R. G. Compton, F. Marken, P. A. Z. Suarez, C. S. Consorti, R. F. de Souza, J. Dupont: "Water-induced accelerated ion diffusion: voltammetric studies in 1-methyl-3-[2,6-(S)-dimethylocten-2-yl]imidazolium tetrafluoroborate, 1-butyl-3-methylimidazolium tetrafluoroborate and hexafluorophosphate ionic liquids", *New J. Chem.* **24** (2000) 1009 – 1015, **DOI:** 10.1039/b007172m.
56. D. R. MacFarlane, P. Meakin, J. Sun, N. Amini, M. Forsyth: "Pyrrolidinium Imides: A New Family of Molten Salts and Conductive Plastic Crystal Phases", *J. Phys. Chem. B* **103** (1999) 4164 – 4170, **DOI**: 10.1021/jp984145s.
57. C. M. Gordon: "Synthesis of Ionic Liquids" in P. Wasserscheid, T. Welton (eds.), *Ionic Liquids in Synthesis*, Wiley-VCH, Weinheim 2003, pp. 7 – 21.
58. P. Walden, "Über die Molekulargrösse und elektrische Leitfähigkeit einiger geschmolzenen Salze", *Bull. Acad. Imper. Sci.* (St. Petersburg) 1914, 405 – 422.
59. J. S. Wilkes, J. A. Levisky, R. A. Wilson, C. L. Hussey: "Dialkylimidazolium chloroaluminate melts: a new class of room-temperature ionic liquids for electrochemistry, spectroscopy and synthesis", *Inorg. Chem.* **21** (1982) 1263 – 1264.
60. R. T. Carlin, J. S. Wilkes: "Chemistry and speciation in room-temperature chloroaluminate molten salts", in G. Mamantov, A. I. Popov (eds.): *Chemistry of Nonaqueous Solutions*, VCH, Weinheim 1994, pp. 277 – 306.
61. C. L. Hussey: "Room temperature haloaluminate ionic liquids. Novel solvents for transition metal solution chemistry", *Pure Appl. Chem.* **60** (1988) 1763 – 1772.
62. H. L. Chum, V. R. Koch, L. L. Miller, R. A. Osteryoung: "Electrochemical scrutiny of organometallic iron complexes and hexamethylbenzene in a room temperature molten salt", *J. Am. Chem. Soc.* **97** (1975) 3264 – 3265.
63. J. Robinson, R. A. Osteryoung: "An electrochemical and spectroscopic study of some aromatic hydrocarbons in the room temperature molten salt system aluminum chloride-n-butylpyridinium chloride", *J. Am. Chem. Soc.* **101** (1979) 323 – 327.
64. H. A. Øye, M. Jagtoyen, T. Oksefjell, J. S. Wilkes: "Vapor pressure and thermodynamics of the system 1-methyl-3-ethyl-imidazolum chloride-aluminum chloride", *Mater. Sci. Forum (Molten Salt Chem. Technol.)* **73-75** (1991) 183 – 189.
65. Y. Chauvin, S. Einloft, H. Olivier: "Catalytic Dimerization of Propene by Nickel-Phosphine Complexes in 1-Butyl-3-methylimidazolium Chloride/AlEtxCl3-x (x = 0, 1) Ionic Liquids", *Ind. Eng. Chem. Res.*, **34** (1995) 1149 – 1155, **DOI**: 10.1021/ie00043a017.
66. B. Gilbert, Y. Chauvin, H. Olivier, F. Di Marco-Van Tiggelen: "Disproportionation of polynuclear chloroethylaluminate anions in acidic 1-butyl-3-methylimidazolium chloride – $AlEtCl_2$ molten salts in the presence of a hydrocarbon phase", *J. Chem. Soc., Dalton Trans.* 1995, no. 23, 3867 – 3871.
67. S. D. Williams, J. P. Schoebrechts, J. C. Selkirk, G. Mamantov: "A new room temperature molten salt solvent system: organic cation tetrachloroborates". *J. Am. Chem. Soc.* **109** (1987) 2218 – 2219, **DOI**: 10.1021/ja00241a069.

68. Y. Chauvin, H. Olivier-Bourbigou: "Nonaqueous ionic liquids as reaction solvents", *CHEMTECH* **25** (1995) 26 – 30.
69. G. W. Parshall: "Catalysis in molten salt media", *J. Am. Chem. Soc.*, **94** (1972) 8716 – 8719.
70. S. I. Lall, D. Mancheno, S. Castro, V. Behaj, J. I. Cohen, R. Engel: "Polycations. Part X. LIPs, a new category of room temperature ionic liquid based on polyammonium salts", *Chem. Commun.* (2000) 2413 – 2414, **DOI:** 10.1039/b007443h.
71. Mitsubishi Petrochemical Co, Nisshin Steel Co Ltd, EP 0404179, 1990 (S. Mori, K. Ida, H. Suzuki, S. Takahashi, I. Saeki).
72. P. B. Hitchcock, K. R. Seddon, T. Welton: "Hydrogen-bond acceptor abilities of tetrachlorometalate(II) complexes in ionic liquids", *J. Chem. Soc., Dalton Trans.* 1993, 2639 – 2643, **DOI:** 10.1039/DT9930002639.
73. P. B. Hitchcock, R. J. Lewis, T. Welton: "Vanadyl complexes in ambient-temperature ionic liquids. The first x-ray crystal structure of a tetrachlorooxovanadate(IV) salt", *Polyhedron* **12** (1993) no. 16, 2039 – 2044, **DOI:** 10.1016/S0277-5387(00)81478-3.
74. A. P. Abbott, G. Capper, D. L. Davies, H. L. Munro, R. K. Rasheed, V. Tambyrajah: "Preparation of novel, moisture-stable, Lewis-acidic ionic liquids containing quaternary ammonium salts with functional side chains", *Chem. Commun.* 2001, no. 19, 2010 – 2011, **DOI:** 10.1039/b106357j.
75. B. K. M. Chan, N. Chang, M. R. Grimmett: "The synthesis and thermolysis of imidazole quaternary salts", *Aust. J. Chem.* **30** (1977) no. 9, 2005 – 2013, **DOI**: 10.1071/CH9772005.
76. F. H. Hurley, T. P. Wier: "Electrodeposition of metals from fused quaternary ammonium salts", *J. Electrochem. Soc.* **98** (1951) 203 – 206.
77. P. J. Dyson, M. C. Grossel, N. Srinivasan, T. Vine, T. Welton, D. J. Williams, A. J. P. White, T. Zigras: "Organometallic synthesis in ambient temperature chloroaluminate(III) ionic liquids. Ligand exchange reactions of ferrocene", *J. Chem. Soc., Dalton Trans.* 1997, 3465 – 3469, **DOI:** 10.1039/a702978k.
78. J. Fuller, R. T. Carlin, H. C. De Long, D. Haworth: "Structure of 1-ethyl-3-methylimidazolium hexafluorophosphate: model for room temperature molten salts", *J. Chem. Soc., Chem. Commun.* 1994, 299 – 300, **DOI:** 10.1039/C39940000299.
79. L. Cammarata, S. G. Kazarian, P. A. Salter, T. Welton: "Molecular states of water in room temperature ionic liquids", *Phys. Chem. Chem. Phys.* **3** (2001) 5192 – 5200, **DOI:** 10.1039/b106900d.
80. N. L. Lancaster, T. Welton, G. B. Young: "A study of halide nucleophilicity in ionic liquids", *J. Chem. Soc., Perkin Trans.* **2** 2001, 2267 – 2270, **DOI:** 10.1039/b107381h.
81. J. Fuller, R. T. Carlin: "Facile preparation of tetrafluoroborate and trifluoromethanesulfonate room-temperature ionic liquids", *Proc. -Electrochem. Soc.* **98** (1999) 227 – 230.
82. A. S. Larsen, J. D. Holbrey, F. S. Tham, C. A. Reed: "Designing Ionic Liquids: Imidazolium Melts with Inert Carborane Anions", *J. Am. Chem. Soc.* **122** (2000) no. 30, 7264 – 7272, **DOI**: 10.1021/ja0007511.
83. D. R. MacFarlane, J. Golding, S. Forsyth, M. Forsyth, G. B. Deacon: "Low viscosity ionic liquids based on organic salts of the dicyanamide anion", *Chem. Commun.* 2001, 1430 – 1431, **DOI:** 10.1039/b103064g.
84. M. Hasan, I. V. Kozhevnikov, M. R. H. Siddiqui, A. Steiner, N. Winterton: "Gold Compounds as Ionic Liquids. Synthesis, Structures, and Thermal Properties of *N*,*N*′-Dialkylimidazolium Tetrachloroaurate Salts", *Inorg. Chem.* **38** (1999) no. 25, 5637 – 5641, **DOI**: 10.1021/ic990657p.
85. C. G. Swain, A. Ohno, D. K. Roe, R. Brown, T. Maugh: "Tetrahexylammonium benzoate, a liquid salt at 25 °C, a solvent for kinetics or electrochemistry", *J. Am. Chem. Soc.* **89** (1967) no. 11, 2648 – 2649, **DOI**: 10.1021/ja00987a025.
86. R. M. Pomaville, S. K. Poole, LaDonna J. Davis, C. F. Poole: "Solute -solvent interactions in tetra-*n*-butylphosphonium salts studied by gas chromatography", *J. Chromatog. A* **438** (1988) 1 – 14.
87. H. Matsumoto, T. Matsuda, Y. Miyazaki: "Room Temperature Molten Salts Based on Trialkylsulfonium Cations and Bis(trifluoromethylsulfonyl)imide", *Chem. Lett.* **29** (2000) no. 12, 1430 – 1431.
88. R. P. Swatloski, J. D. Holbrey, R. D. Rogers: "Ionic liquids are not always green: hydrolysis of 1-butyl-3-methylimidazolium hexafluoroborate", *Green Chem.* **5** (2003) no. 4, 361 – 363.
89. V. R. Koch, C. Nanjundiah, G. Battista Appetecchi, B. Scrosati: "The interfacial

stability of Li with two new solvent-free ionic liquids: 1,2-dimethyl-3-propylimidazolium imide and methide", *J. Electrochem. Soc.* **142** (1995) no. 7, L116 – L118.
90. C. F. Poole, B. R. Kersten, S. S. J. Ho, M. E. Coddens, K. G. Furton: "Organic salts, liquid at room temperature, as mobile phases in liquid chromatography", *J. Chromatogr. A* **352** (1986) 407 – 425, **DOI**: 10.1016/S0021-9673(01)83397-5
91. Y. Chauvin, L. Mussmann, H. Olivier: "A Novel Class of Versatile Solvents for Two-Phase Catalysis: Hydrogenation, Isomerization, and Hydroformylation of Alkenes Catalyzed by Rhodium Complexes in Liquid 1,3-Dialkylimidazolium Salts", *Angew. Chem. Int. Ed. Engl.* **34** (1995) 2698 – 2700.
92. C. L. Hussey: "Room temperature molten salt systems", *Adv. Molten Salt Chem.* **5** (1983) 185 – 230.
93. W. T. Ford, R. J. Hauri, D. J. Hart: "Syntheses and properties of molten tetraalkylammonium tetraalkylborides", *J. Org. Chem.* **38** (1973) no. 22, 3916 – 3918, **DO**I: 10.1021/jo00962a020.
94. R. Damico: "Preparation, Characterization, and Reactions of Lithium and Sodium Tetraalkylboron Compounds", *J. Org. Chem.* **29** (1964) 1971 – 1976, **DOI**: 10.1021/jo01030a077.
95. S. K. Poole, P. H. Shetty, C. F. Poole: "Chromatographic and spectroscopic studies of the solvent properties of a new series of room-temperature liquid tetraalkylammonium sulfonates", *Anal. Chim. Acta* **218** (1989) 241 – 264, **DOI:** 10.1016/S0003-2670(00)80302-5.
96. S. A. Bolkan, J. T. Yoke: "Room temperature fused salts based on copper(I) chloride-1-methyl-3-ethylimidazolium chloride mixtures. 1. Physical properties", *J. Chem. Eng. Data* **31** (1986) no. 2, 194 – 197, **DOI**: 10.1021/je00044a019.
97. S. A. Bolkan, J. T. Yoke: "Room-temperature fused salts based on copper(I) chloride-1-methyl-3-ethylimidazolium chloride mixtures. 2. Reactions with dioxygen", *Inorg. Chem.* **25** (1986) no. 20, 3587 – 3590, **DOI**: 10.1021/ic00240a011.
98. J. D. Holbrey, W. M. Reichert, R. P. Swatloski, G. A. Broker, W. R. Pitner, K. R. Seddon, R. D. Rogers: "Efficient, halide free synthesis of new, low cost ionic liquids: 1,3-dialkylimidazolium salts containing methyl-and ethyl-sulfate anions", *Green Chem.* **4** (2002) 407 – 413, **DOI:** 10.1039/b204469b.
99. J. F. Knifton, "Ethylene-glycol from synthesis gas *via* ruthenium melt catalysis", *J. Am. Chem. Soc.* **103** (1981) no. 13, 3959 – 3961.
100. J. F. Knifton: "Syngas reactions. XI. The ruthenium melt catalyzed oxonation of internal olefins", *J. Mol. Catal.* **43** (1987) no. 1, 65 – 78, **DOI**: 10.1016/0304-5102(87)87022-0.
101. J. F. Knifton: "Syngas reactions. XIII. The ruthenium melt-catalyzed oxonation of terminal olefins", *J. Mol. Catal.* **47** (1988) no. 1, 99 – 116, **DOI**: 10.1016/0304-5102(88)85077-6.
102. C. J. Bradaric, A. Downard, C. Kennedy, A. J. Robertson, Y. H. Zhou: "Industrial preparation of phosphonium ionic liquids", in R. D. Rogers, K. R. Seddon (eds.): *Ionic Liquids as Green Solvents: Progress and Prospects*, ACS Symp. Ser., Vol. 856, American Chemical Society, Washington D.C. 2003, pp. 41 – 56.
103. C. J. Bradaric, A. Downard, C. Kennedy, A. J. Robertson, Y. H. Zhou: "Industrial preparation of phosphonium ionic liquids", *Green Chem.* **5** (2003) no. 2, 143 – 152.
104. J. McNulty, A. Capretta, J. Wilson, J. Dyck, G. Adjabeng, A. Robertson: "Suzuki cross-coupling reactions of aryl halides in phosphonium salt ionic liquid under mild conditions", *Chem. Commun.* 2002, no. 17, 1986 – 1987.
105. Cytec Tech. Corp., WO 0187900, 2001 (A. J. Robertson).
106. Cytec Tech. Corp., WO 02 079212, 2002 (A. J. Robertson, K. R. Seddon).
107. Cytec Tech. Corp., WO 03 020843, 2003 (A. J. Robertson, M. J. Earle, C. Hardacre, K. R. Seddon, A. Ramani).
108. Rhône Poulenc Chimie, WO 2002092608 2002 (M. Garayt, V. Le Boulaire, D. Gree, R. Gree, V. Schanen, J.-F. Spindler).
109. Rhône Poulenc Chimie, FR 2839716, 2003 (V. Schanen, R. Gree, D. Gree, M. Garayt, V. Le Boulaire).
110. Univ. of Leicester, WO 0056700, 2000 (A. P. Abbott, D. L. Davies).
111. M. Deetlefs, K. R. Seddon: "Improved preparations of ionic liquids using microwave irradiation", *Green Chem.* **5** (2003) no. 2, 181 – 186.
112. N. E. Leadbeater, H. M. Torenius, H. Tye: "Ionic liquids as reagents and solvents in conjunction with microwave heating: rapid synthesis of alkyl halides from alcohols and nitriles from aryl halides", *Tetrahedron* **59** (2003) 2253 – 2258.

113. J. M. Lévêque, G. Cravotto: "Microwaves, power ultrasound, and ionic liquids. A new synergy in green organic synthesis", *CHIMIA* **60** (2006) 313 – 320.
114. V. V. Namboodiri, R. S. Varma: "An improved preparation of 1,3-dialkylimidazolium tetrafluoroborate ionic liquids using microwaves", *Tetrahedron Lett.* **43** (2002) no. 31, 5381 – 5383, **DOI**: 10.1016/S0040-4039(02)01075-4.
115. S.-K. Fu, S.-T. Liu: "Preparation of functionalized imidazolium salts under microwave irradiation", *Synth. Commun.* **36** (2006) 2059 – 2067.
116. J. Hoffmann, M. Nüchter, B. Ondruschka, P. Wasserscheid: "Ionic liquids and their heating behaviour during microwave irradiation – a state of the art report and challenge to assessment", *Green Chem.* **5** (2003) 296 – 299, **DOI:** 10.1039/b212533a.
117. BASF AG, WO 2005070896, 2005 (M. Maase, K. Massonne, S. Laszlo).
118. Covalent Associates Inc., WO 02094883, 2002 (R. X. Ren, V. R. Koch).
119. Cytec Canada Inc., WO 2004094438, 2004 (C. J. Bradaric-Baus, Y. Zhou).
120. Merck, WO 2004106288, 2004 (U. Welz-Biermann, N. M. Ignatyev, H. Willner, G. Bissky).
121. Merck, WO 2005049555, 2005 (N. M. Ignatyev, U. Welz-Biermann, H. Willner, A. Kucheryna).
122. Merck, WO 2006010455, 2006 (N. M. Ignatyev, U. Welz-Biermann, M. Finze, E. Bernhardt).
123. Sachem Inc., WO 2004005222, 2004 (R. Moulton).
124. Sachem Inc., WO 2004016571, 2004 (R. Moulton).
125. Scionix Ltd, US 6573405, 2003 (A. Abbott, D. L. Davies).
126. Solvent Innovation GmbH, WO 03074494, 2003 (P. Wasserscheid, A. Bösmann, R. van Hal).
127. Solvent Innovation GmbH, WO 2004035542, 2004 (P. Wasserscheid, C. Steffens, B. Driessen-Hölscher, C. Hilgers).
128. Solvent Innovation GmbH, EP 1622877, 2006 (P. Wasserscheid, R. van Hal, C. Hilgers)
129. M. J. Earle, K. R. Seddon: "Ionic liquids. Green solvents for the future", *Pure Appl. Chem.* **72** (2000) no. 7, 1391 – 1398.
130. M. Earle, K. R. Seddon: "Ionic liquids: Green solvents for the future", in M. Abraham, L. Moens (eds.): *Clean Solvents: Alternative Media for Chemical Reactions and Processing*, ACS Symposium Series, Vol. 819, American Chemical Society, Washington D.C. 2002, pp. 10 – 25.
131. M. J. Earle: "Ionic liquids: Solvents for the twenty-first century", in R. D. Rogers, K. R. Seddon 8eds.): *Ionic Liquids: Industrial Applications for Green Chemistry*, ACS Symp. Ser., Vol. 818, American Chemical Society, Washington D.C. 2002, pp. 90 – 105.
132. J. S. Wilkes: "Properties of ionic liquid solvents for catalysis" *J. Mol. Catal. A: Chem.* **214** (2004), no. 1, 11 – 17, **DOI:** 10.1016/j.molcata.2003.11.029.
133. H. Olivier-Bourbigou, L. Magna: "Ionic liquids: perspectives for organic and catalytic reactions", *J. Mol. Catal. A: Chem.* **182-183** (2002) 419 – 437, **DOI:** 10.1016/S1381-1169(01)00465-4.
134. N. L. Lancaster, T. Welton, G. B. Young: "A study of halide nucleophilicity in ionic liquids", *J. Chem. Soc., Perkin Trans.* **2**, 2001, 2267 – 2270, **DOI**: 10.1039/b107381h.
135. N. L. Lancaster, P. A. Salter, T. Welton, G. B. Young: "Nucleophilicity in Ionic Liquids. 2.[1] Cation Effects on Halide Nucleophilicity in a Series of Bis(trifluoromethylsulfonyl)imide Ionic Liquids", *J. Org. Chem.* **67** (2002) 8855 – 8861, **DOI**: 10.1021/jo026113d.
136. R. Sheldon: "Catalytic reactions in ionic liquids", *Chem. Commun.* 2001, no. 23, 2399 – 2407, **DOI:** 10.1039/b107270f.
137. J. Dupont, R. F. de Souza, P. A. Z. Suarez: "Ionic Liquid (Molten Salt) Phase Organometallic Catalysis", *Chem. Rev*. **102** (2002) 3667 – 3692, **DOI:** 10.1021/cr010338r.
138. D. B. Zhao, M. Wu, Y. Kou, E. Min: "Ionic liquids: applications in catalysis", *Catal. Today* **74** (2002) no. 1-2, 157 – 189, **DOI:** 10.1016/S0920-5861(01)00541-7..
139. C. E. Song: "Enantioselective chemo-and bio-catalysis in ionic liquids", *Chem. Commun.* 2004, no. 9, 1033 – 1043. **DOI:** 10.1039/b309027b.
140. P. J. Dyson: "Review: Synthesis of organometallics and catalytic hydrogenations in ionic liquids", *Appl. Organomet. Chem.* **16** (2002) no. 9, 495 – 500, **DOI:** 10.1002/aoc.337.
141. H. Olivier, Y. Chauvin: "Non-aqueous Room-Temperature Ionic Liquids: A New Class of Solvents for Catalytic Organic Reactions", *Proc. Electrochem. Soc.* **96** (1996) no. 7, 70 – 73.

142. Queen's Univ. of Belfast, WO 03 028883, 2003 (K. R. Seddon, C. Hardacre, B. J. McAuley).
143. Queen's Univ. of Belfast, WO 02 072519, 2002 (M. J. Earle, B. J. McAuley, A. Ramani, K. R. Seddon, J. M. Thomson).
144. S. Csihony, H. Mehdi, Z. Homonnay, A. Vertes, O. Farkas, I. T. Horvath, "In situ spectroscopic studies related to the mechanism of the Friedel-Crafts acetylation of benzene in ionic liquids using $AlCl_3$ and $FeCl_3$", *J. Chem. Soc., Dalton Trans.* **5** (2002) 680 – 685, **DOI:** 10.1039/b109303g.
145. ICI Plc, WO 0132308, 2001 (M. H. Valkenberg, E. Sauvage, C. P. DeCastro-Moreira, W. F. Hölderich).
146. M. H. Valkenberg, C. De Castro, W. F. Hölderich: "Immobilisation of ionic liquids on solid supports", *Green Chem*, **4** (2002) 88 – 93, **DOI:** 10.1039/b107946h.
147. C. J. Adams, M. J. Earle, G. Roberts, K. R. Seddon: "Friedel-Crafts reactions in room temperature ionic liquids", *Chem. Commun.* 1998, no. 19, 2097 – 2098, **DOI:** 10.1039/a805599h.
148. J. A. Boon, J. A. Levisky, J. L. Pflug, J. S. Wilkes: "Friedel-Crafts reactions in ambient-temperature molten salts", *J. Org. Chem.* **51** (1986) no. 4, 480 – 483, **DOI**: 10.1021/jo00354a013.
149. M. J. Earle, U. Hakala, B. J. McAuley, M. Nieuwenhuyzen, A. Ramani, K. R. Seddon: "Metal bis {(trifluoromethyl)sulfonyl}amide complexes: highly efficient Friedel – Crafts acylation catalysts", *Chem. Commun.* (2004) no. 12, 1368 – 1369, **DOI:** 10.1039/b403650f.
150. J. S. Wilkes: "Friedel-Crafts reaction in chloroaluminate molten salts", in G. Mamantov, R. Marrasi (eds.): Molten Salt Chemistry, An Introduction and Selected Applications, *NATO ASI Ser., Ser. C, 202*, D. Reichel Publishing, Dordrecht 1987, pp. 405 – 416.
151. P. Wasserscheid, M. Sesing, W. Korth: Hydrogensulfate and tetrakis(hydrogensulfato)borate ionic liquids: synthesis and catalytic application in highly Brønsted-acidic systems for Friedel – Crafts alkylations" *Green Chem.*, **4** (2002) 134 – 138, **DOI:** 10.1039/b109845b.
152. J. Ross, J. Xiao: "Friedel – Crafts acylation reactions using metal triflates in ionic liquid", *Green Chem.* **4** (2002) 129 – 133, **DOI:** 10.1039/b109847k.
153. S. Kobayashi, K. A. Jørgensen (eds.): *Cycloaddition Reactions in Organic Synthesis*, Wiley-VCH, Weinheim 2002.
154. D. E. Kaufmann, M. Nouroozian, H. Henze: "Molten Salts as an Efficient Medium for Palladium Catalyzed C-C Coupling Reactions", *Synlett* 1996, no. 11, 1091 – 1092, **DOI**: 10.1055/s-1996-5658.
155. A. J. Carmichael, M. J. Earle, J. D. Holbrey, P. B. McCormac, K. R. Seddon: "The Heck reaction in ionic liquids: a multiphasic catalyst system", *Org. Lett.* **1** (1999) no. 7, 997 – 1000, **DOI**: 10.1021/ol9907771.
156. I. P. Beletskaya, A. V. Cheprakov: "The Heck Reaction as a Sharpening Stone of Palladium Catalysis", *Chem. Rev.* **100** (2000) no. 8, 3009 – 3066, **DOI**: 10.1021/cr9903048.
157. K. R. Seddon, A. Stark: "Selective catalytic oxidation of benzyl alcohol and alkylbenzenes in ionic liquids", *Green Chem.* **4** (2002) no. 2, 119 – 123, **DOI:** 10.1039/b111160b.
158. C. J. Adams, M. J. Earle, K. R. Seddon: "Catalytic cracking reactions of polyethylene to light alkanes in ionic liquids", *Green Chem.* **2** (2000) 21 – 24, **DOI:** 10.1039/a908167d.
159. Queen's Univ. of Belfast, WO 02 094740, 2002 (K. Anderson, P. Goodrich, C. Hardacre, S. E. J. McCath).
160. P. J. Dyson, D. J. Ellis, D. G. Parker, T. Welton: "Arene hydrogenation in a room-temperature ionic liquid using a ruthenium cluster catalyst", *Chem. Commun.* 1999, no. 1, 25 – 26, **DOI:** 10.1039/a807447j.
161. C. J. Adams, M. J. Earle, K. R. Seddon: "Stereoselective hydrogenation reactions in chloroaluminate(III) ionic liquids: a new method for the reduction of aromatic compounds", *Chem. Commun.* 1999, no. 11, 1043 – 1044, **DOI:** 10.1039/a901302d.
162. A. J. Carmichael, D. M. Haddleton, S. A. F. Bon, K. R. Seddon: "Copper(I) mediated living radical polymerisation in an ionic liquid", *Chem. Commun.* 2000, no. 114, 1237 – 1238, **DOI:** 10.1039/b003335i.
163. A. J. Carmichael, D. M. Haddleton: "Polymer synthesis in ionic liquids", in P. Wasserscheid, T. Welton (eds.): *Ionic Liquids in Synthesis*, Wiley-VCH, Weinheim 2003, pp. 319 – 335.
164. C. Hardacre, J. D. Holbrey, S. P. Katdare, K. R. Seddon: "Alternating copolymerisation of styrene and carbon monoxide in ionic liquids", *Green Chem.* **4** (2002) 143 – 146, **DOI:** 10.1039/b111157b.
165. P. Kubisa: "Application of ionic liquids as solvents for polymerization processes", *Prog.*

Polym. Sci. **29** (2004) 3 – 12, **DOI**: 10.1016/j.progpolymsci.2003.10.002.
166. ICI Plc, WO 0037400, 2000 (N. Winterton, K. R. Seddon, Y. Patell).
167. M. J. Earle, P. B. McCormac, K. R. Seddon: "Diels – Alder reactions in ionic liquids. A safe recyclable alternative to lithium perchlorate – diethyl ether mixtures", *Green Chem.* **1** (1999) 23 – 25, **DOI:** 10.1039/a808052f.
168. A. J. Walker, N. C. Bruce: "Cofactor-dependent enzyme catalysis in functionalized ionic solvents", *Chem. Commun.* **22** (2004) 2570 – 2571.
169. R. A. Sheldon, R. M. Lau, M. J. Sorgedrager, F. van Rantwijk, K. R. Seddon: "Biocatalysis in ionic liquids", *Green Chem.* **4** (2002) 147 – 151, **DOI:** 10.1039/b110008b.
170. F. van Rantwijk, R. M. Lau, R. A. Sheldon, "Biocatalytic transformations in ionic liquids", *Trends Biotechnol.* **21** (2003) 131 – 138, **DOI:** 10.1016/S0167-7799(03)00008-8.
171. S. G. Cull, J. D. Holbrey, V. Vargas-Mora, K. R. Seddon, G. J. Lye: "Room-temperature ionic liquids as replacements for organic solvents in multiphase bioprocess operations", *Biotechnol. Bioeng.* **69** (2000) no. 2, 227 – 233.
172. R. M. Lau, F. van Rantwijk, K. R. Seddon, R. A. Sheldon: "Lipase-Catalyzed Reactions in Ionic Liquids", *Org. Lett.* **2** (2000) no. 26, 4189 – 4191.
173. M. Erbeldinger, A. J. Mesiano, A. J. Russell: "Enzymatic catalysis of formation of Z-aspartame in ionic liquid -An alternative to enzymatic catalysis in organic solvents", *Biotechnol. Prog.* **16** (2000) no. 6, 1129 – 1131.
174. R. A. Sheldon, F. van Rantwijk, R. M. Lau: "Biotransformations in Ionic Liquids: An Overview" in R. D. Rogers, K. R. Seddon (eds.): *Ionic Liquids as Green Solvents: Progress and Prospects*, ACS Symposium Series 856, American Chemical Society, Washington DC 2003, pp. 192 – 205.
175. S. Park, R. J. Kazlauskas: "Biocatalysis in ionic liquids – advantages beyond green technology", *Curr. Opin. Biotech.* **14** (2003) no. 4, 432 – 437, **DOI**: 10.1016/S0958-1669(03)00100-9.
176. U. Kragl, M. Eckstein, N. Kaftzik: "Enzyme catalysis in ionic liquids", *Curr. Opin. Biotech.* **13** (2002) no. 6, 565 – 571, **DOI:** 10.1016/S0958-1669(02)00353-1.
177. M.-Y. Lee, J. S. Dordick: "Enzyme activation for nonaqueous media", *Curr. Opin. Biotech.* **13** (2002) no. 4, 376 – 384, **DOI:** 10.1016/S0958-1669(02)00337-3.
178. H. S. Schöfer, N. Kaftzik, P. Wasserscheid, U. Kragl: "Enzyme catalysis in ionic liquids: lipase catalysed kinetic resolution of 1-phenylethanol with improved enantioselectivity", *Chem. Commun.* (2001) no. 5, 425 – 426, **DOI**: 10.1039/b009389k.
179. T. Itoh, E. Akasaki, K. Kudo, S. Shirakami: "Lipase-Catalyzed Enantioselective Acylation in the Ionic Liquid Solvent System: Reaction of Enzyme Anchored to the Solvent", *Chem. Lett.* **30** (2001) no. 3, 262 – 263.
180. K. W. Kim, B. Song, M. Y. Choi, M. J. Kim: "Biocatalysis in Ionic Liquids: Markedly Enhanced Enantioselectivity of Lipase", *Org. Lett.* **3** (2001) no. 10, 1507 – 1509, **DOI**: 10.1021/ol015824f.
181. J. D. Holbrey: "Industrial Applications of Ionic Liquids", *Chim. Oggi* **22** (2004) no. 6, 35 – 37.
182. B. Cornils, W. A. Herrmann: "Concepts in homogeneous catalysis: the industrial view", *J. Catal.* **216** (2003) no. 1-2, 23 – 31, **DOI:** 10.1016/S0021-9517(02)00128-8.
183. D. J. Cole-Hamilton: "Homogeneous Catalysis – New Approaches to Catalyst Seperation, Recovery, and Recycling", *Science* **299** (2003) no. 5613 1702 – 1706.
184. B. Driessen-Hölscher: "Multiphase homogeneous catalysis", *Adv. Catal.* **42** (1998) 473 – 505.
185. D. E. Kaufmann, M. Nouroozian, H. Henze: "Molten Salts as an Efficient Medium for Palladium Catalyzed C-C Coupling Reactions", *Synlett* 1996, no. 11, 1091 – 1092, **DOI:** 10.1055/s-1996-5658.
186. W. A. Herrmann, K. Öfele, D. von Preysing, S. K. Schneider: "Phospha-palladacycles and N-heterocyclic carbene palladium complexes: efficient catalysts for CC-coupling reactions", *J. Organomet. Chem.* **687** (2003) no. 2, 229 – 248, **DOI:** 10.1016/j.jorganchem.2003.07.028.
187. A. Corma, H. Garcia: "Lewis Acids: From Conventional Homogeneous to Green Homogeneous and Heterogeneous Catalysis", *Chem. Rev.* **103** (2003) no. 11, 4307 – 4366, **DOI**: 10.1021/cr030680z.
188. BP Chem. Int. Ltd., WO 9521806, 1995 (A. K. Abdul-Sada et al.).
189. Chevron Chem. Co., US 6,395,948, 2000 (K. D. Hope, M. S. Driver, T. V. Harris).
190. D. Commereuc, Y. Chauvin, G. Leger, J. Gaillard, "The Dimersol process for

dimerisation of olefins – a chemical approach", *Revue de L'Institut Français du Petrole*, **37** (1982) 639 – 649.
191. Y. Chauvin, B. Gilbert, I. Guibard: "Catalytic dimerization of alkenes by nickel complexes in organochloroaluminate molten salt", *J. Chem. Soc., Chem. Commun.* 1990, 1715 – 1716, **DOI:** 10.1039/C39900001715.
192. C. Tiyapiboonchaiya, D. R. MacFarlane, J. Sun, M. Forsyth: "Polymer-in-ionic-liquid electrolytes", *Macromol. Chem. Phys.* **203** (2003) no. 13, 1906 – 1911.
193. C. L. Hussey, T. B. Scheffler, J. S. Wilkes, A. Fannin Jr.: "Chloroaluminate Equilibria in the Aluminum Chloride-1-Methyl-3-ethylimidazolium Chloride Ionic Liquid", *J. Electrochem. Soc.* **133** (1986) no. 7, 1389 – 1391.
194. T. Ito, T. Nohira: "Non-conventional electrolytes for electrochemical applications", *Electrochim. Acta* **45** (2000), no. 15-16, 2611 – 2622, doi:10.1016/S0013-4686(00)00341-8.
195. R. T. Carlin, H. C. De Long, J. Fuller, P. C. Trulove: "Dual Intercalating Molten Electrolyte Batteries", *J. Electrochem. Soc.* **141** (1994) no. 7, L73 – L76.
196. D. R. MacFarlane, J. Huang, M. Forsyth: "Lithium-doped plastic crystal electrolytes exhibiting fast ion conduction for secondary batteries", *Nature* **402** (1999) no. 6763, 792 – 794.
197. A. Lewandowski, A. Swiderska: "Electrochemical capacitors with polymer electrolytes based on ionic liquids", *Solid State Ionics* **161** (2003) no. 3-4, 243 – 249, **DOI:** 10.1016/S0167-2738(03)00275-3.
198. A. B. McEwen, S. F. McDevitt, V. R. Koch: "Nonaqueous Electrolytes for Electrochemical Capacitors: Imidazolium Cations and Inorganic Fluorides with Organic Carbonates", *J. Electrochem. Soc.* **144** (1997) no. 4, L84 – L86.
199. M. Ue, K. Ida, S. Mori: "Electrochemical Properties of Organic Liquid Electrolytes Based on Quaternary Onium Salts for Electrical Double-Layer Capacitors", *J. Electrochem. Soc.* **141** (1994) no. 11, 2989 – 2996.
200. M. Grätzel: "Dye-sensitized solar cells", *J. Photochem. Photobiol. C: Photochem. Rev*. **4** (2003) 145 – 153.
201. H. Matsumoto, T. Matsuda: "Wet type solar cells using ionic liquid as electrolyte (2)", *Electrochem.* **70** (2002) no. 3, 190 – 194.
202. L. M. Goldenberg, R. A. Osteryoung: "Benzene polymerization in 1-ethyl-3-methylimidazolium chloride-AlCl3 ionic liquid", *Synth. Met.* **64** (1994) no. 1, 63 – 68, doi:10.1016/0379-6779(94)90276-3.
203. X.-H. Xu, C. L. Hussey: "Electrodeposition of Silver on Metallic and Nonmetallic Electrodes from the Acidic Aluminum Chloride-1-Methyl-3-Ethylimidazolium Chloride Molten Salt", *J. Electrochem. Soc.* **139** (1992) no. 5, 1295 – 1300.
204. G. E. McManiss III, A. N. Fletcher, D. E. Bliss, US 4624755, 1986.
205. Y. G. Zhao, T. J. VanderNoot: "Electrodeposition of aluminium from nonaqueous organic electrolytic systems and room temperature molten salts", *Electrochim. Acta* **42** (1997) no. 1, 3 – 13, **DOI:** 10.1016/0013-4686(96)00080-1.
206. F. Endres: "Ionic Liquids: Solvents for the Electrodeposition of Metals and Semiconductors", *ChemPhysChem*. **3** (2002) no. 2, 144 – 154.
207. J. Liu, G. B. Jiang, Y. G. Chi, Y. Q. Cai, Q. X. Zhou, J.T. Hu: "Use of ionic liquids for liquid-phase microextraction of polycyclic aromatic hydrocarbons", *Anal. Chem.* **75** (2003), no. 21, 5870 – 5876.
208. G. A. Baker, S. N. Baker, S. Pandrey, F. Bright: "An analytical view of ionic liquids", *The Analyst* **130** (2005) 800 – 808.
209. F. Mutelet, V. Butet, J.-N. Jaubert: "Application of Inverse Gas Chromatography and Regular Solution Theory for Characterization of Ionic Liquids", *Ind. Eng. Chem. Res.* **44** (2005) no. 11, 4120 – 4127, **DOI:** 10.1021/ie048806l.
210. Jing-fu Liu, Gui-bin Jiang, Jing-fu Liu, J. Å. Jönsson: "Application of ionic liquids in analytical chemistry", *TrAC, Trends Anal. Chem.* **24** (2005) 20 – 27, **DOI:** 10.1016/j.trac.2004.09.005.
211. J. L. Anderson, D. W. Armstrong, G.-T. Wei: "Ionic Liquids in Analytical Chemistry", *Anal. Chem.* **78** (2006) 2892 – 2902.
212. S. Pandey: "Analytical applications of room-temperature ionic liquids: A review of recent efforts", *Anal. Chim. Acta* **556** (2006) 38 – 45, **DOI:** 10.1016/j.aca.2005.06.038.
213. A. M. Stalcup, B. Cabovska: "Ionic liquids in chromatography and capillary electrophoresis", *J. Liq. Chromatogr. Relat. Technol.* **27** (2004) no. 7-9, 1443 – 1459.
214. A. E. Visser: "Metal ion separations in aqueous biphasic systems and room

temperature ionic liquids", PhD thesis, University of Alabama, Tuscaloosa, AL. 2002.
215. A. E. Visser, R. P. Swatloski, S. T. Griffin, D. H. Hartman, R. D. Rogers: "Liquid/liquid extraction of metal ions in Room Temperature Ionic Liquids", *Sep. Sci. & Techn.* **36** (2001) no. 5-6, 785 – 804.
216. A. E. Visser et al.: "Task-specific ionic liquids for the extraction of metal ions from aqueous solutions", *Chem. Commun.* 2001, no. 1, 135 – 136.
217. G.-T. Wei, Z. Yang, C.-J. Chen: "Room temperature ionic liquid as a novel medium for liquid/liquid extraction of metal ions", *Anal. Chim. Acta* **488** (2003) no. 2, 183 – 192.
218. M. L. Dietz: "Ionic liquids as extraction solvents: Where do we stand?", *Sep. Sci. Technol.* **41** (2006), (10) 2047 – 2063, **DOI:** 10.1080/01496390600743144.
219. A. G. Fadeev, M. M. Meagher: "Opportunities for ionic liquids in recovery of biofuels", *Chem. Commun.* 2001, no. 3, 295 – 296.
220. A. Heintz, J. K. Lehmann, C. Wertz: "Thermodynamic Properties of Mixtures Containing Ionic Liquids. 3. Liquid-Liquid Equilibria of Binary Mixtures of 1-Ethyl-3-methylimidazolium Bis(trifluoromethylsulfonyl)imide with Propan-1-ol, Butan-1-ol, and Pentan-1-ol", *J. Chem. Eng. Data* **48** (2003) no. 3, 472 – 474.
221. C.-T. Wu, K. N. Marsh, A. V. Deev, J. A. Boxal: "Liquid-Liquid Equilibria of Room-Temperature Ionic Liquids and Butan-1-ol", *J. Chem. Eng. Data* **48** (2003) no. 3, 486 – 491.
222. T. M. Letcher, N. Deenadayalu, B. Soko, D. Ramjugernath, P. K. Naicker: "Ternary Liquid-Liquid Equilibria for Mixtures of 1-Methyl-3-octylimidazolium Chloride + an Alkanol + an Alkane at 298.2 K and 1 bar", *J. Chem. Eng. Data* **48** (2003) no. 4, 904 – 907.
223. T. M. Letcher, P. Reddy: "Ternary Liquid-Liquid Equilibria for Mixtures of 1-Hexyl-3-methylimidozolium (tetrafluoroborate or hexafluorophosphate) + Ethanol + an Alkene at T=298.2 K", *Fluid Phase Equilib.* **219** (2004) no. 2, 107 – 112.
224. A. Bösmann, L. Datsevich, A. Jess, A. Lauter, C. Schmitz, P. Wasserscheid: "Deep desulfurization of diesel fuel by extraction with ionic liquids", *Chem. Commun.* 2001, 2494 – 2495.
225. S. G. Zhang, Z. C. Zhang: "Novel properties of ionic liquids in selective sulfur removal from fuels at room temperature", *Green Chem.* **4** (2002) no. 4, 376 – 379.
226. S. Zhang, Q. Zhang, Z. C. Zhang: "Extractive Desulfurization and Denitrogenation of Fuels Using Ionic Liquids", *Ind. Eng. Chem. Res.* **43** (2004) no. 2, 614 – 622.
227. J. Eßer, P. Wasserscheid, A. Jess: "Deep desulfurization of oil refinery streams by extraction with ionic liquids", *Green Chem.* **6** (2004) no. 7, 316 – 322.
228. A. Arce, O. Rodríguez, A. Soto: "Experimental Determination of Liquid-Liquid Equilibrium Using Ionic Liquids: *tert*-Amyl Ethyl Ether + Ethanol + 1-Octyl-3-Methylimidazolium Chloride System at 298.15 K", *J. Chem. Eng. Data* **49** (2004) no. 3, 514 – 517.
229. A. Arce, O. Rodríguez, A. Soto: "*tert*-Amyl Ethyl Ether Separation from Its Mixtures with Ethanol Using the 1-Butyl-3-methylimidazolium Trifluoromethanesulfonate Ionic Liquid: Liquid-Liquid Equilibrium", *Ind. Eng. Chem. Res.* **43** (2004) no. 26, 8323 – 8327.
230. L. C. Branco, J. G. Crespo, C. A. M. Afonso: "Highly Selective Transport of Organic Compounds by Using Supported Liquid Membranes Based on Ionic Liquids", *Angew. Chem. Int. Ed.* **41** (2002) no. 15, 2771 – 2773.
231. L. C. Branco, J. G. Crespo, C. A. M. Afonso: "Studies on the Selective Transport of Organic Compounds by Using Ionic Liquids as Novel Supported Liquid Membranes", *Chem. Eur. J.* **8** (2002) no. 17, 3865 – 3871.
232. M. Matsumoto, Y. Inomoto, K. Kondo: "Selective separation of aromatic hydrocarbons through supported liquid membranes based on ionic liquids", *J. Membr. Sci.* **246** (2005) no. 1, 77 – 81.
233. J.-F. Liu, Y.-G. Chi, G.-B Jiang: "Screening the extractability of some typical environmental pollutants by ionic liquids in liquid-phase microextraction", *J. Sep. Sci.* **28** (2005) no. 1, 87 – 91.
234. M. S. Selvan, M. D. McKinley, R. H. Dubois, J. L. Atwood: "Liquid-Liquid Equilibria for Toluene + Heptane + 1-Ethyl-3-methylimidazolium Triiodide and Toluene + Heptane + 1-Butyl-3-methylimidazolium Triiodide", *J. Chem. Eng. Data* **45** (2000) 841 – 845.
235. Carl v. Ossietzky Univ. Oldenburg, 101 54 052 A1, 2003 (J. Gmehling, M. Krummen).
236. T. M. Letcher, N. Deenadayalu: "Ternary liquid-liquid equilibria for mixtures of 1-methyl-3-octyl-imidazolium

chloride+benzene+an alkane at T=298.2K and 1 atm", *J. Chem. Thermodyn.* **35** (2003) no. 1, 67 – 76.
237. T. M. Letcher, P. Reddy: "Ternary (liquid+liquid) equilibria for mixtures of 1-hexyl-3-methylimidazolium (tetrafluoroborate or hexafluoroborate)+benzene+an alkane at T = 298.2 K and p = 0.1 MPa", *J. Chem. Thermodyn.* **37** (2005) 415 – 421.
238. BASF AG, WO 02074718, 2002 (W. Arlt, C. Jork, M. Seiler, T. Schneider).
239. G. W. Meindersma, J. G. Podt, M. Gutiérrez Meseguer, A. B. de Haan: "Ionic liquids as alternatives to organic solvents in liquid-liquid extraction of aromatics", in Robin D. Rogers, Kenneth Seddon (eds.): *Ionic Liquids IIIB, Fundamentals, Progress, Challenges, and Opportunities*, ACS Symposium Series 902, American Chemical Society, Washington DC 2005, pp. 57 – 71.
240. G. W. Meindersma, J. G. Podt, A. B. de Haan: "Selection of ionic liquids for the extraction of aromatic hydrocarbons from aromatic/aliphatic mixtures", *Fuel Process. Technol.* **87** (2005) no. 1, 59 – 70, **doi:**10.1016/j.fuproc.2005.06.002.
241. G. W. Meindersma, J. G. Podt, M. B. Klaren, A. B. de Haan: "Ionic liquids for BTEX separations", International Solvent Extraction Conference, ISEC 2005, Beijing, 19 – 23 September 2005, paper D201, pp. 1308 – 1314.
242. G. W. Meindersma, J. G. Podt, M. B. Klaren, A. B. de Haan: "Separation of aromatic and aliphatic hydrocarbons with ionic liquids", *Chem. Eng. Commun.* **193** (2006) no. 11, 1384 – 139, **DOI**: 10.1080/009864405005114036.
243. G. W. Meindersma, J. G. Podt, A. B. de Haan: "Ternary liquid-liquid equilibria for mixtures of toluene+n-heptane+an ionic liquid", *Fluid Phase Equilib.* **247** (2006) no. 1-2, 158 – 168, **DOI:** 10.1016/j.fluid.2006.07.002.
244. G. W. Meindersma, J. G. Podt, A. B. de Haan: "Ternary liquid-liquid equilibria for mixtures of an aromatic+an aliphatic hydrocarbon+[mebupy]BF_4", *J. Chem. Eng. Data* **51** (2006) no. 5, 1814 – 1819, **DOI**: 10.1021/je060198o.
245. S. N. V. K. Aki, B. R. Mellein, E. M. Saurer, J. F. Brennecke: "High-Pressure Phase Behavior of Carbon Dioxide with Imidazolium-Based Ionic Liquids", *J. Phys. Chem. B* **108** (2004) no. 52, 20355 – 20365, **DOI**: 10.1021/jp046895+.
246. J. L. Anthony, J. L. Anderson, E. J. Maginn, J. F. Brennecke: "Anion Effects on Gas Solubility in Ionic Liquids", *J. Phys. Chem. B* **109** (2005) 6366 – 6374, **DOI**: 10.1021/jp046404l.
247. R. E. Baltus, B. H. Culbertson, S. Dai, H. Luo, D. W. DePaoli: "Low-Pressure Solubility of Carbon Dioxide in Room-Temperature Ionic Liquids Measured with a Quartz Crystal Microbalance", *J. Phys. Chem. B* **108** (2004) no. 2, 721 – 727, **DOI**: 10.1021/jp036051a.
248. P. Scovazzo, D. Camper, J. Kieft, J. Poshusta, C. Koval, R. Noble: "Regular Solution Theory and CO_2 Gas Solubility in Room-Temperature Ionic Liquids", *Ind. Eng. Chem. Res.* **43** (2004) no. 21, 6855 – 6860, **DOI**: 10.1021/ie049601f.
249. M. B. Shiflett, A. Yokozeki: "Solubilities and Diffusivities of Carbon Dioxide in Ionic Liquids: [bmim][PF_6] and [bmim][BF_4]", *Ind. Eng. Chem. Res.* **44** (2005) no. 12, 4453 – 4464, **DOI**: 10.1021/ie058003d.
250. R. E. Baltus, R. M. Counce, B. H. Culbertson, H. Luo, D. W. DePaoli, S. Dai, D. C. Duckworth: "Examination of the Potential of Ionic Liquids for Gas Separations", *Sep. Sci Technol.* **40** (2005) no. 1-3, 525 – 541, **DOI**: 10.1081/SS-200042513.
251. A. Shariati, C. J. Peters: "High-pressure phase equilibria of systems with ionic liquids", *J. Supercrit. Fluids* **34** (2005) no. 2, 171 – 176, doi:10.1016/j.supflu.2004.11.011.
252. E. D. Bates, R. D. Mayton, I. Ntai, J. H. Davis, Jr.: "CO_2 Capture by a Task-Specific Ionic Liquid", *J. Am. Chem. Soc.* **124** (2002) no. 6, 926 – 927, **DOI**: 10.1021/ja017593d.
253. C. Cadena, J. L. Anthony, J. K. Shah, T. I. Morrow, J. F. Brennecke, E. J. Maginn: "Why Is CO_2 So Soluble in Imidazolium-Based Ionic Liquids?", *J. Am. Chem. Soc.* **126** (2004) no. 16, 5300 – 5308, **DOI**: 10.1021/ja039615x.
254. D. Camper, C. Becker, C. Koval, R. Noble: "Low Pressure Hydrocarbon Solubility in Room Temperature Ionic Liquids Containing Imidazolium Rings Interpreted Using Regular Solution Theory", *Ind. Eng. Chem. Res.* **44** (2005) no. 6, 1928 – 1933, **DOI:** 10.1021/ie049312r.
255. D. Camper, C. Becker, C. Koval, R. Noble: "Diffusion and Solubility Measurements in Room Temperature Ionic Liquids", *Ind. Eng. Chem. Res.* **45** (2006) no. 1, 445 – 450, **DOI**: 10.1021/ie0506668.
256. D. Camper, P. Scovazzo, C. Koval, R. Noble: "Gas Solubilities in Room-Temperature Ionic Liquids", *Ind. Eng. Chem. Res.* **43** (2004) no. 12, 3049 – 3054, **DOI**: 10.1021/ie034097k.

257. D. Morgan, L. Ferguson, P. Scovazzo: "Diffusivities of Gases in Room-Temperature Ionic Liquids: Data and Correlations Obtained Using a Lag-Time Technique", *Ind. Eng. Chem. Res.* **44** (2005) no. 13, 4815 – 4823, **DOI**: 10.1021/ie048825v.
258. Chevron, US 6339182, 2002 (C. L. Munson, L. C. Boudreau, M. S. Driver, W. Schinski).
259. Chevron, US 6623659, 2003 (C. L. Munson, L. C. Boudreau, M. S. Driver, W. Schinski).
260. J. D. Holbrey, W. M. Reichert, R. G. Reddy, R. D. Rogers, "Heat Capacities of Ionic Liquids and Their Applications as Thermal Fluids", in R. D. Rogers, K. R. Seddon (eds.): *Ionic Liquids as Green Solvents: Progress and Prospects*; ACS Symposium Series 856; American Chemical Society: Washington DC 2003, pp. 121 – 133.
261. C. Ye, W. Liu, Y. Chen, L. Yu: "Room-temperature ionic liquids: a novel versatile lubricant", *Chem. Commun.* 2001 no. 21, 2244 – 2245, **DOI:** 10.1039/b106935g.
262. W. Liu, C. Ye, Q. Gong, H. Wang, P. Wang: "Tribological Performance of Room-Temperature Ionic Liquids as Lubricant". *Trib. Letts*. **13** (2002) no. 2, 81 – 85, **DOI**: 10.1023/A:1020148514877.
263. W. Liu, C. Ye, Y. Chen, Z. Ou, D.C. Sun: "Tribological behavior of sialon ceramics sliding against steel lubricated by fluorine-containing oils", *Trib. Int.* **35** (2002) no. 8, 503 – 509, **DOI**: 10.1016/S0301-679X(02)00044-0.
264. H. Wang, Q. Lua, C. Yea, W. Liu, Z. Cui: "Friction and wear behaviors of ionic liquid of alkylimidazolium hexafluorophosphates as lubricants for steel/steel contact" *Wear* **256** (2004) no. 1-2, 44 – 48, **DOI:** 10.1016/S0043-1648(03)00255-2.
265. R. A. Reich, P. A. Stewart, J. Bohaychick, J. A. Urbanski: "Base oil properties of ionic liquids", *Lubr. Eng.* **59** (2003) no. 7, 16 – 21.
266. M. Freemantle: "BASF's Smart Ionic Liquid", *Chem. Eng. News* **81** (2003) no. 13, 9.
267. K. R. Seddon: "A taste of the future", *Nature Mat.* **2** (2003) no. 1, 363 – 364.
268. BASF AG, WO 03062251 (A1), WO03062171 (A3), 2003 (M. Volland et al.).
269. M. Maase: "Erstes technisches Verfahren mit ionischen Flüssigkeiten", *Chem. Unserer Zeit* **38** (2004) 434 – 436.
270. R. D. Rogers, K. R. Seddon: "Ionic Liquids — Solvents of the Future?", *Science* **302** (2003) 792 – 793.
271. G. Parkinson: "Ionic Liquids Make An Environmental Splash", *Chem. Eng. Prog.* **100** (2004) no. 9, 7 – 9.
272. M. Maase, K. Masonne: "Biphasic Acid Scavenging Utilizing Ionic Liquids: The First Commercial Process With Ionic Lqiuids", in R. D. Rogers, K. S. Seddon (eds.): *Ionic Liquids IIIB: Fundamentals, Progress, Challenges, and Opportunities. Transformations and Processes*, ACS Symposium Series 902, American Chemical Society, Washington DC 2005, pp. 126 – 132.
273. BASF AG, WO 2005061416, 2005 (M. Maase, O. Huttenloch).
274. S. N. Falling, S. A. Godleski, J. R. Monnier, G. W. Phillips, J. S. Kanel: "Development of an Industrial Process for the Lewis Acid/Iodide-Salt Catalyzed Rearrangement of 3,4-Epoxy-1-butene to 2,5-Dihydrofuran", *Book of Abstracts, 1st International Congress on Ionic Liquids (COIL)*, Salzburg, Austria 19-22 June 2005, pp. 58 – 59.
275. Eastman Kodak Co., US 5238889, 1993 (S. N. Falling, S. A. Godleski, L. W. MCGarry, T. R. Nolen, J. S. Kanel).
276. Eastman Chem. Co., US 5315019, 1994 (G. W. Philips, S. N. Falling, S. A. Godleski, J. R. Monnier).
277. J. S. Kanel, S. N. Falling, G. W. Phillips, J. R. Monnier, S. A. Godleski: "Extractive Catalyst Recovery in an Ionic Liquid Process for 2,5-Dihydrofuran" AIChE Annual Meeting 2006 San Francisco, USA, Extractive Separations 02C00, 2006.
278. BASF AG, WO 2005026089, 2005 (V. Stegmann, K. Masonne).
279. C. R. Schmid, C. A. Beck, J. S. Cronin, M. A. Staszak: "Demethylation of 4-Methoxyphenylbutyric Acid Using Molten Pyridinium Hydrochloride on Multikilogram Scale", *Org. Proc. Res. Dev.* **8** (2004) no. 4, 670 – 673, **DOI**: 10.1021/op0499526 S1083-6160(04)09952-9.
280. J. F. W. McOmie, M. L. Watts, D. E. West: "Demethylation of aryl methyl ethers by boron tribromide" *Tetrahedron* **24** (1968) 2289 – 2292.
(b) Boron Tribromide, in L. Paquette (ed.): *Electronic Encyclopedia of Reagents for Organic Synthesis*; John Wiley & Sons: New York, 1995.
281. D. Papa, E. Schwenk, H. Hankin: "Friedel-Crafts Synthesis of ω-Aroyl Aliphatic Acids", *J. Am. Chem. Soc.* **69** (1947) 3018 – 3022.

282. C. S. Yi, L. C. Martinelli, C. D. Blanton, Jr.: "Synthesis of N-methyl-1-oxa-5-aza[10]paracyclophane: a conformationally restricted analog of phenoxypropylamines", *J. Org. Chem.* **43** (1978) 405 – 409, **DOI**: 10.1021/jo00397a006.
283. B. Weyershausen, K. Hell, U. Hesse: "Industrial application of ionic liquids as process aid", *Green Chem.* **7** (2005) no. 5, 283 – 287.
284. Goldschmidt, EP 1382630, 2004 (K. Hell, U. Hesse, B. Weyershausen).
285. Arkema/Atofina. WO 0181353, 2001 (Ph. Bonnet, E. Lacroix, J.-P. Schirmann).
286. IFP, US 5723712, 1998 (Y. Chauvin, R. De Souza, H. Olivier).
287. H. Olivier-Bourbigou, A. Forestière: "Ionic liquids in multiphasic reactions", in P. Wasserscheid, T. Welton (eds.): *Ionic Liquids in Synthesis*, Wiley-VCH, Weinheim 2003, pp. 258 – 280.
288. F. Favre, A. Forestière, F. Hugues, H. Olivier-Bourbigou, J. A. Chodorge: "Butenes dimerisation: From monophasic Dimersol(tm) to biphasic Difasol(tm)", *Oil Gas European Magazine* **31** (2005) 83 – 87.
289. Chevron Phillips US 6 395 948, 2002 (K. D. Hope, M. S. Driver, T. V. Harris).
290. C. Jork, M. Seiler, Y.-A. Beste, W. Arlt: "Influence of Ionic Liquids on the Phase Behavior of Aqueous Azeotropic Systems", *J. Chem. Eng. Data* **49** (2004) no. 4, 852 – 857.
291. BASF, WO 05/016484, 2005 (Y. A. Beste, H. Schoenmakers, W. Arlt, M. Seiler, C. Jork).
292. G. W. Meindersma, G. S. Vos, M. B. Klaren, A. B. de Haan: "Evaluation of an RDC for extraction with ionic liquids", Proceedings 1st International Congress on Ionic Liquids (COIL), Salzburg, A., 19 – 22 June 2005, Dechema, Frankfurt am Main 2005, p. 45.
293. G. W. Meindersma, G. S. Vos, M. B. Klaren, A. B. de Haan: "Evaluation of contactor performance for extraction with ionic liquids", International Solvent Extraction Conference, ISEC 2005, Beijing, 19 – 23 September 2005, paper B301, pp. 735 – 741.
294. G. W. Meindersma, J. G. Podt, G. S. Vos, A. B. de Haan: Selection of ionic liquids for BTEX separation and evaluation of extraction in an RDC", Abstracts of the Congress Topics Achema 2006. Green Solvents for Environmentally Friendly Processes, Frankfurt am Main, D., 15 – 19 May, 2006, Paper 1130, p. 103.
295. A. P. Abbott, G. Capper, D. L. Davies, R. K. Rasheed: "Ionic Liquid Analogues Formed from Hydrated Metal Salts", *Chem. Eur. J.* **10** (2004) 3769 – 3774.
296. A. P. Abbott, G. Capper, D. L. Davies, R. K. Rasheed, J. Archer, C. John: "Electrodeposition of chromium black from ionic liquids", *Trans. Inst. Met. Finish.* **82** (2004) no. 1-2, 14 – 17.
297. A. P. Abbott, G. Capper, D. L. Davies, R. K. Rasheed, V. Tambyrajah: "Novel solvent properties of choline chloride/urea mixtures", *Chem. Commun.* 2003, no. 1, 70 – 71, **DOI**: 10.1039/b210714g.
298. A. P. Abbott, D. Boothby, G. Capper, D. L. Davies, R. K. Rasheed: "Deep Eutectic Solvents Formed between Choline Chloride and Carboxylic Acids: Versatile Alternatives to Ionic Liquids", *J. Am. Chem. Soc.* **126** (2004) no. 29, 9142 – 9147, **DOI**: 10.1021/ja048266j.
299. A. P. Abbott, G. Capper, B. G. Swain, D. A. Wheeler: "Electropolishing of stainless steel in an ionic liquid", *Trans. Inst. Met. Finish.* **83** (2005) no. 1, 51 – 53.
300. B. Weyershausen, K. Lehmann: "Industrial application of ionic liquids as performance additives", *Green Chem.* **7** (2005) no. 1, 15 – 19, **DOI**: 10.1039/b411357h.
301. Goldschmidt, EP 1566413, 2005 (K. Lehmann, S. Silber, B. Weyershausen).
302. A. Bösmann, T. Schubert: "Identification of industrial applications for Ionic Liquids: High-Performance-Additives for the use in Hi-Tech-Cleaning-Solutions", Poster "Green Solvents For Synthesis", 3 – 6 October 2004, Bruchsal, Germany, http://www.iolitec.de/download/poster/-Poster%20Bruchsal%20V2.pdf.
303. Olin Corp., US 4 744 21, 1988 (K. O. Knollmueller).
304. Air Products and Chemicals, US 2004206241, 2004 (D. J. Tempel, Ph. B. Henderson, J. R. Brzozowski).
305. Air Products and Chemicals. US 2006060818, 2006 (D. J. Tempel, Ph. B. Henderson, J. R. Brzozowski, R. M. Pearlstein, D. Garg).
306. M. Kömpf, "Mobility under high pressure" *Linde Technology*, (2006) (January), 24-29, http://www.linde.com/international/web/linde/-likelindeeng.nsf/docbyalias/nav_technology.
307. R. Giernoth, D. Bankmann, N. Schlörer: "High performance NMR in ionic liquids", *Green Chem.* **7** (2005) 279 – 282, **DOI:** 10.1039/b417783e.

308. S. Busi, M. Lahtinen, M. Kärnä, J. Valkonen, E. Kolehmainen, K. Rissanen: "Synthesis, characterization and thermal properties of nine quaternary dialkyldiaralkylammonium chlorides", *J. Mol. Struct.* **787** (2006) no. 1-3, 18 – 30, **DOI**: 10.1016/j.molstruc.2005.10.027.
309. A. J. Carmichael, C. Hardacre, J. D. Holbrey, M. Nieuwenhuyzen, K. R. Seddon: "A Method for Studying the Structure of Low-Temperature Ionic Liquids by XAFS", *Anal. Chem.* **71** (1999) 4572 – 4574, **DOI**: 10.1021/ac990332q.
310. S. Caporali, U. Bardi, A. Lavacchi: "X-ray photoelectron spectroscopy and low energy ion scattering studies on 1-buthyl-3-methyl-imidazolium bis(trifluoromethane) sulfonamide", *J. Electron Spectrosc. Relat. Phenom.* **151** (2006) no. 1, 4 – 8, **DOI**: 10.1016/j.elspec.2005.09.010.
311. C. Pretti, C. Chiappe, D. Pieraccini, M. Gregori, F. Abramo, G. Monni, L. Intorre: "Acute toxicity of ionic liquids to the zebrafish (Danio rerio)", *Green Chem.* **8** (2005) no. 3, 238 – 240, **DOI:** 10.1039/b511554j.
312. A. S. Wells, V. T. Coombe: "On the freshwater ecotoxicity and biodegradation properties of some common ionic liquids", *Org. Process Res. Dev.* **10** (2006) no. 10, 794 – 798.
313. K. M. Docherty, C. F. Kulpa jr: "Toxicity and antimicrobial activity of imidazolium and pyridinium ionic liquids", *Green Chem*. **7** (2005) no. 4, 185 – 189**, DOI**: 10.1039/b419172b.
314. R. J. Bernot, E. E. Kennedy, G. A. Lamberti: "Environmental Toxicology -Effects of Ionic Liquids on the Survival, Movement, and Feeding Behavior of the Freshwater Snail, Physa acuta", *Environ. Toxicol. Chem.* **24** (2005) no. 7, 1759 – 1765
315. J. Pernak, K. Sobaszkiewic, I. Mirksa: "Anti-microbial activities of ionic liquids". *Green Chem.* **5** (2003) 52 – 56.
316. J. Pernak, I. Goc, I. Mirska: "Anti-microbial activities of protic ionic liquids with lactate anion", *Green Chem.* **6** (2004) 323 – 329.
317. R. J. Bernot, M. A. Brueseke, M. A. Evans-White, G. A. Lamberti: "Environmental Toxicology -Acute and Chronic Toxicity of Imidazolium-Based Ionic Liquids on *Daphnia magna*", *Environ. Toxicol. Chem.* **24** (2005) no. 1, 87 – 92.
318. S. Stolte, J. Arning, U. Bottin-Weber, M. Matzke, F. Stock, K. Thiele, M. Uerdingen, U. Welz-Biermann, B. Jastorff, J. Ranke: "Anion effects on the cytotoxicity of ionic liquids", *Green Chem.* **8** (2006) 621 – 629, **DOI**: 10.1039/b602161a.
319. D. J. Couling, R. J. Bernot, K. M. Docherty, J.-N. K. Dixon, E. J. Maginn: "Assessing the factors responsible for ionic liquid toxicity to aquatic organisms *via* quantitative structure – property relationship modeling", *Green Chem.* **8** (2006) 82 – 90, **DOI:** 10.1039/b511333d.
320. E. A. Hassoun, M. Abraham, V. Kini, M. Al-Ghafri, A. Abushaban: "Cytotoxicity of the ionic liquid, 1-N-butyl-3-methyl imidazolium chloride", *Res. Commun. Pharmacol. Toxicol.* **7** (2002) 23 – 31.
321. R. P. Swatloski, J. D. Holbrey, S. B. Memon, G. A. Caldwell, K. A. Caldwell, R. D. Rogers: "Using *Caenorhabditis elegans* to probe toxicity of 1-alky-3-methylimidiazolium chloride based ionic liquids", *Chem. Commun.* 2004, 668 – 669.
322. A. Latala, P. Stepnowski, M. Nedzia, W. Mrozik: "Marine toxicity assessment of imidazolium ionic liquids: Acute effects on the Baltic algae *Oocystis submarina* and *Cyclotella meneghiniana*", *Aquat. Toxicol.* **73** (2005) no. 1, 91 – 98, **DOI**: 10.1016/j.aquatox.2005.03.008.
323. N. Gathergood, M. T. Garcia, P. J. Scammells: "Biodegradable ionic liquids: Part I. Concept, preliminary targets and evaluation". *Green Chem.* **6** (2004) 166 – 175.
324. M. T. Garcia, N. Gathergood, P. J. Scammells: "Biodegradable ionic liquids. Part II. Effect of the anion and toxicology", *Green Chem.* **7** (2005) no. 1, 9 – 14, **DOI:** 10.1039/b411922c.
325. N. Gathergood, P. J. Scammells, M. T. Garcia: "Biodegradable ionic liquids. Part III. The first readily biodegradable ionic liquids", *Green Chem.* **8** (2006) 156 – 160.
326. M. Matsumoto, K. Mochiduki, K. Fukunishi, K. Kondo: "Extraction of organic acids using imidazolium-based ionic liquids and their toxicity to *Lactobacillus rhamnosus*", *Sep. Pur. Technol.* **40** (2004) no. 1, 97 – 101, **DOI**: 10.1016/j.seppur.2004.01.009.
327. M. Matsumoto, K. Mochiduki, K. Kondo: "Toxicity of ionic liquids and organic solvents to lactic acid-producing bacteria", *J. Biosc. Bioeng.* **98** (2004) no. 5, 344 – 347, **DOI**: 10.1016/S1389-1723(04)00293-2.
328. F. Stock, J. Hoffmann, J. Ranke, R. Störmann, B. Ondruschka, B. Jastorff: "Effects of ionic liquids on the acetylcholinesterase – a

structure-activity relationship consideration", *Green Chem.* **6** (2004) 286 – 290.

329. B. Jastorff et al.: "How sustainable are ionic liquids? Structure-activity relationships and biological testing as important elements for sustainability evaluation", *Green Chem.* **5** (2003) 136 – 142.
330. J. Ranke et al.: "Biological effects of imidazolium ionic liquids in acute *Vibrio fischeri* and WST1 cell viability assays". *Ecotoxicol. Environ. Saf.* **28** (2004) no. 3, 396 – 404, **DOI**: 10.1016/S0147-6513(03)00105-2.
331. B. Jastorff et al.: "Progress in evaluation of risk potential of ionic liquids — basis for an eco-design of sustainable products", *Green Chem.* **7** (2005) no. 5, 362 – 372, **DOI**: 10.1039/b418518h.
332. K. M. Docherty, S. Z. Hebbeler, C. F. Kulpa Jr.: "An assessment of ionic liquid mutagenicity using the Ames Test", *Green Chem.* **8** (2006) no. 6, 560 – 567, **DOI**: 10.1039/b602418a.
333. M. Uerdingen, C. Treber, M. Balser, G. Schmitt, C. Werner: "Corrosion behaviour of ionic liquids", *Green Chem.* **7** (2005) no. 5, 321 – 325, **DOI**: 10.1039/b419320m.
334. I. Perissi, U. Bardi, S. Caporali, A. Lavacchi: "High temperature corrosion properties of ionic liquids", *Corros. Sci.* **48** (2006) no. 9, 2349 – 2362, **DOI**: 10.1016/j.corsci.2006.06.010.

Nuclear Magnetic Resonance Imaging

SIEGFRIED STAPF, Department of Technical Physics II, Technische Universität Ilmenau, 98684 Ilmenau, Germany

SONG-I HAN, Department of Chemistry and Biochemistry, University of California, Santa Barbara, CA, CA 93106-9510, USA

1. Introduction 1
2. Principles of Spatial Encoding in NMR 2
2.1. Repetition of NMR Facts 2
2.2. k Space 5
2.3. Phase Encoding and Frequency Encoding 7
2.4. Slice Selection 9
2.5. 2D and 3D Imaging 10
2.6. Spatial Resolution 12
3. Image Contrast 13
3.1. Relaxation 13
3.2. Diffusion 15
3.3. Chemical Shift Imaging 17
3.4. Spin Density and Parameter Images . 19
4. Fast Imaging Techniques 20
4.1. Multislicing Methods 21
4.2. Nonequilibrium Methods 21
4.3. Multiple Spin-Echo Sequences 23
4.4. Multiple Gradient Echo Sequences . . 24
5. Measurements of Transport Processes 25
5.1. q Space 25
5.2. Motion Encoding and Moments 28
5.3. Velocity Maps 31
5.4. Time-of-Flight Methods 33
6. Hardware and Materials 36
6.1. Imaging and Micro-Imaging Hardware 36
6.2. Imaging X Nuclei 37
6.3. Liquids, Solids, and Gases 38
7. References 38

Nuclear magnetic resonance imaging plays an increasingly important role in the investigation of processes and reactions. It introduces a wide range of contrast parameters that make it suitable for combining the well-established spectroscopic performance of NMR spectroscopy with schemes for obtaining spatially resolved data. Clinical magnetic resonance imaging (MRI) is already a standard diagnostic tool with ever-increasing capabilities, more flexible detection schemes, and fast acquisition being developed. Nonclinical MRI can benefit from this development, but it is required to solve quite different problems with the same basic methodology: noninvasive process control, reaction monitoring, materials studies, transport phenomena, and quality control are some of the topics encountered in industrial chemistry. This article summarizes the necessary background of NMR imaging and the parameters that usually enter into the image information or can be extracted from it at the control of the experimenter. In particular, the use and effect of pulsed field gradients for encoding position and motion is discussed. A number of fast imaging methods are presented which are employed for routine use, and focus is put on the monitoring of transport processes, which are of particular interest to the field. Finally, some essential guidelines regarding the use of particular hardware are given, and the strategies for imaging liquid, solid, and gaseous substances are described.

1. Introduction

Nuclear magnetic resonance imaging and industrial chemistry — an unusual combination? It is true that NMR imaging has not entered the factory environment on a routine basis yet. What most nonspecialists know about NMR can be summarized by two different applications that have the physical principle in common, but rely on different and highly specialized commercial hardware. The first is NMR spectroscopy, which indeed is used routinely not only in academia and research centers, but also in a wide range of companies, with the goal of following synthesis processes or to perform quality control on chemical substances in off-line or on-line mode. Secondly, imaging techniques [commonly known

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.l17_l01

as magnetic resonance imaging (MRI) or MR tomography] have found their most potent commercial application in medical scanners that are available in most major hospitals nowadays, and have been employed for an ever-increasing number of specific applications developing out of traditional brain and spine monitoring towards routine investigations of joints, soft tissue, and lungs.

Because clinical MRI is optimized for the particular problem of imaging large objects (i.e., humans) with medium spatial resolution (1–3 mm)within an acceptable timescale (typically $<$ 30 min), it is of limited usefulness for other applications; besides, the high degree of specialization has led to rather cost-intensive hardware solutions. However, the same physical principles that govern the imaging of tissue can be used for objects of arbitrary shape and size between micrometers and meters. The development of hardware that allows one to perform imaging protocols with superconducting magnets and spectrometers very similar to those used for spectroscopy have widened the field during the last decade to include materials studies and transport phenomena. This laboratory-based technology, often called NMR microscopy (indeed, it provides submillimeter resolution but is normally not on a par with the performance of optical microscopy), has been used to investigate complex structures by means of detecting the NMR signal of solid, liquid, or gas phases, and to monitor transport and reaction within these systems. They are suitable for fundamental studies on real-scale or purpose-built devices. More recent approaches attempt to provide in situ capabilities, e.g., quality control or process control at the production site itself. Because of the need for portability and flexibility, they often rely on lightweight and, consequently, low magnetic field equipment which is subject to loss of sensitivity and imaging capabilities. However, a number of promising attempts have been made to exploit the principle of NMR in unconventional environments and under suboptimal conditions (see Chapter 6).

While MRI provides solutions to many problems in industrial chemistry, the proper combination of methods and analysis is required, and some conditions regarding the feasibility and resolution of MRI can be given even a priori. This article summarizes the necessary background of NMR imaging and the parameters that usually enter into the image information or can be extracted from it at the control of the experimenter; only the very fundamentals of NMR itself are given, and the reader is otherwise referred to → Nuclear Magnetic Resonance and Electron Spin Resonance Spectroscopy. It then provides a list of "fast" imaging methods (i.e., those that pass the test of routine use under realistic conditions) and puts a focus on the monitoring of transport processes, which are of particular interest to the field. Finally, some essential guidelines regarding the use of particular hardware and the feasibility of measuring certain substances are given.

2. Principles of Spatial Encoding in NMR

2.1. Repetition of NMR Facts

To understand and discuss NMR imaging, one can mostly avoid a full description of the origin of the NMR signal and its theoretical derivation, which would involve delving deeply into quantum mechanics. The argument for a substantial simplification of the problem lies to a certain degree in the statistical nature of the NMR effect itself: while the underlying resonance phenomena appear on a nuclear level and therefore must follow the particular rules of quantum mechanics, the output of an imaging experiment results from an averaging over a large number of spins or molecules, so that one can treat the system like an array of macroscopic magnetizations, such as a set of compass needles. Note, however, that certain fundamental effects, such as those due to mutual coupling of spins, are not reproduced in this framework, and the classical description must always be taken cum grano salis. For a deeper insight into the subject, the reader is referred to standard textbooks covering the theoretical background and general aspects [1 – 6], or in-depth discussions of NMR spectroscopy [7 – 13]. A broad, dictionary-style treatment of all topics can be found in [14, 15].

The principle of NMR itself, which is necessary to understand its extension to NMR imaging, can be summarized in just a few essential equations that describe the magnitude and time

dependence of the acquired signal. First of all, how large is the signal amplitude? When the spins of an arbitrary sample become polarized in a strong magnetic field $\boldsymbol{B}_0$, a net magnetization results which is parallel or antiparallel to $\boldsymbol{B}_0$. The value of this magnetization $\boldsymbol{M}_0$ is not the quantity usually measured; in fact, it is possible to determine it by detecting the force on a balance with the sample inside and outside of the magnetic field, but this datum is of very limited use. By far the majority of the information obtained in NMR stems from the time dependence of the magnetization which, once removed from equilibrium — for instance by rotating it into a direction perpendicular to the magnetic field, hereby employing a pulse of rf irradiation at the resonance frequency — precesses about $\boldsymbol{B}_0$ and returns to its equilibrium due to processes called *relaxation*. The precessing magnetization induces an a.c. voltage inside the receiver coil, wound perpendicular to $\boldsymbol{B}_0$. This voltage $U_{\mathrm{ind}}(t)$ is recorded, and both the frequency spectrum and the phase of the signal are extracted via Fourier transformations. The voltage is proportional to the time derivative of the magnetization

$$U_{\mathrm{ind}}(t) \propto \frac{\mathrm{d}\boldsymbol{M}(t)}{\mathrm{d}t} = \gamma\, \boldsymbol{M}(t) \times \boldsymbol{B}(t) \tag{1}$$

where subscripts have been removed to indicate that neither $\boldsymbol{B}$ nor $\boldsymbol{M}$ is, in general, an equilibrium and time-independent quantity, and γ is the gyromagnetic ratio specific for a given isotope. Neglecting relaxation for the moment, which leads to the eventual damping of the signal and describes the return to the spin system's equilibrium, the transverse (observable) fraction of the magnetization $\boldsymbol{M}_+(t)$ can be written as

$$\boldsymbol{M}_+(t) = \boldsymbol{M}_0(t) \exp(i\omega_0 t) \tag{2}$$

so that its time-derivative is proportional to the initial magnitude M_0 as well as to the precession frequency ω_0:

$$U_{\mathrm{ind}} \propto \frac{\mathrm{d}\boldsymbol{M}_+(t)}{\mathrm{d}t} \propto \boldsymbol{M}_0 \omega_0 \tag{3}$$

The first factor, the amplitude M_0, is given by Curie's law:

$$\boldsymbol{M}_0 = N \frac{\gamma^2 \hbar^2 I(I+1)}{3 k_{\mathrm{B}} T} \boldsymbol{B}_0, \tag{4}$$

where I is the nuclear spin quantum number (see below), $\hbar$ Planck's constant, k_{B} Boltzmann's constant, N the number of nuclei, and T the temperature. To maximize signal intensity, and therefore magnetization, one would thus want to work at a high magnetic field, use a sufficiently large sample, and detect nuclei with a large gyromagnetic ratio γ. Of all stable isotopes, the nucleus of the hydrogen atom ^{1}H has the largest γ. Furthermore, it is contained in all organic matter, and the ^{1}H isotope has almost 100 % natural abundance, which is the reason why the majority of imaging experiments is carried out with ^{1}H nuclei. However, imaging experiments have also been performed successfully on other isotopes, called "X nuclei" in the NMR language.

Curie's law describes a macroscopic magnetization, but its origin is found in the nuclear equivalent of the magnetization, the *spin*. This quantity resembles a compass needle which is located at the core of the nucleus; brought into a magnetic field, its energy state will depend on its orientation relative to the field direction. The spin is a quantum property, i.e., it can only assume quantized (half-integer or integer) values, given by the total spin quantum number: $I = 0$, 1/2, 1, ...

From this picture, the origin of both Curie's law and the resonance phenomenon can be derived. First, nuclei with different energy states will populate the respective levels in a statistical manner, obeying the Boltzmann distribution, which for two spin states +1/2, −1/2, reads:

$$\frac{n_{-\frac{1}{2}}}{n_{+\frac{1}{2}}} = \exp\left(-\frac{\hbar\omega_0}{k_{\mathrm{B}} T}\right) \tag{5}$$

If one inserts typical values for the magnetic field strength (say, 10 T) and room temperature (T = 298 K), the population difference turns out to be only a tiny fraction of 0.007 % for ^{1}H nuclei. This small bias is what provides the equilibrium value M_0.

In the above equation, the energy difference between the two levels is already expressed by its frequency equivalent $\Delta E = \hbar\omega_0$, where ω_0 is the Larmor frequency that denotes, for instance, the frequency of the rf irradiation that is used to induce transitions between levels and thus to manipulate the magnetization. At the same time, it describes the precession frequency of the magnetization about the direction of $\boldsymbol{B}_0$, giving rise to the induced voltage in the detector coil, as was mentioned above. The spin energy levels, and thus the Larmor frequency, are proportional

to the magnetic field strength, which is written in the simplest but most fundamental equation for NMR imaging:

$$\omega = \gamma |\boldsymbol{B}| \tag{6}$$

Because $\boldsymbol{B}$ is not just the externally applied field, but the magnetic field at the local position of the nucleus itself, which may vary from one nucleus to another, a wealth of information is included in the value of ω, or rather its spectral distribution. A large part of the toolbox of NMR is built around this simple dependence. Two regimes can be distinguished that provide entirely different information about the system:

The *microscopic regime* is given by the immediate vicinity of the nucleus. It is surrounded by electrons the motion of which — just like the motion of any electric charge — induces a magnetic field that shields the nucleus from the external field, resulting in the nucleus' specific local magnetic field. The single electron of the hydrogen atom shields a fraction of some 10^{-5} of the external field. Because the shielding depends on the actual charge distribution, which, in turn, is a consequence of the molecular environment of the hydrogen atom, a particular moiety can be identified by the shielding effect it has on the nucleus. From $\omega = \gamma \boldsymbol{B}$ we see that the resonance frequency of all nuclei varies proportionally to the field strength. The difference relative to a standard sample is called the *chemical shift* and is measured in unitless numbers, given as ppm (parts per million, 10^{-6}). Comparing all proton-containing chemical substances, the total range of ^{1}H nuclei resonance frequencies covers about 12 ppm. For instance, in an external magnetic field of 9.4 T, i.e., at a resonance frequency of 400 MHz, the maximum difference in frequencies observed is only about 4800 Hz. To distinguish subtle differences between molecules, the resolution of a good spectrometer must be much better; often values of 10^{-10} (i.e., 0.4 Hz in this example) can be reached (see → Nuclear Magnetic Resonance and Electron Spin Resonance Spectroscopy).

The *macroscopic regime* is the one which is directly accessible by the magnet designer. By introducing, on purpose, an inhomogeneity into the magnetic field by means of additional coils, $\boldsymbol{B}$ and therefore ω are made functions of the position. Of course, this only makes sense if each value of the field $\boldsymbol{B}$ occurs only once in the entire sample volume, so that an unambiguous assignment of the position is possible. The most obvious solution is the generation of a linear field dependence. The relation between position and detected signal is introduced in Section 2.1.

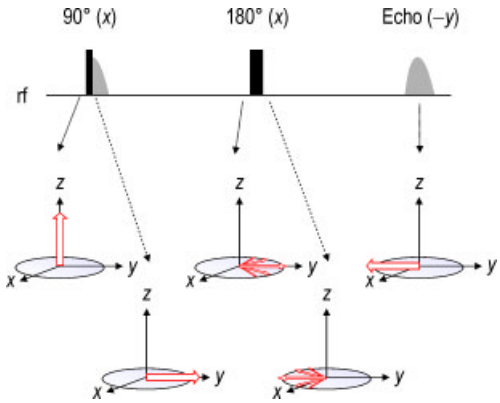

Figure 1. Behavior of the magnetization in a simple echo experiment. Top: a free induction decay (FID) follows the first 90° pulse; x denotes the phase of the pulse, i.e., the axis about which the magnetization is effectively rotated. The 180° pulse is applied with the same phase; the echo appears at twice the separation between the two pulses and its phase is inverted to that of the initial FID. Bottom: the magnetization vector at five stages of the sequence drawn in a coordinate frame rotating at ω^0 about the z axis: before the 90° pulse, the magnetization is in equilibrium, i.e., parallel to the magnetic field (z); immediately after the 90° pulse, it has been rotated (by 90°) into the transverse (xy) plane; as it is composed of contributions from different parts of the sample, the slightly different local fields lead to a loss of coherence, i.e., to a free induction decay in the pick-up coil; before the 180° pulse, the different contributions (narrower arrows) have lost part or all of their coherence; after the 180° pulse, each partial magnetization has been flipped (by 180°) due to the effect of the electromagnetic field of the pulse, but still sees the same, slightly different local fields; at a time corresponding to twice the pulse spacing, the different phase shifts relative to the average value have been reverted, all partial magnetizations are in phase again, and their signal contributions add up coherently — a spin echo is generated.

However, before spatial encoding is described, the basic building block of an imaging experiment is explained by a simplified graphic representation of a basic NMR *pulse sequence*. The effect of either the microscopic or the macroscopic regime on the magnetization are visualized by the schematic picture given in Figure 1 that relates the behavior of the magnetization (the broad arrows in the sketches at the bottom) with the events that take place in

the transmitter/receiver unit of the spectrometer (top row). In the beginning, the *longitudinal magnetization* is in equilibrium and no signal is detected. The first rf pulse (called a 90° pulse since the pulse duration is just long enough to rotate the longitudinal magnetization by that angle) delivers $\boldsymbol{M}$ to the transverse plane and creates *transverse magnetization*. Here, the x in the subscript of 90° indicates a rotation about the x axis to place the transverse magnetization along the y direction. Immediately after such a 90° pulse, the magnetization still has its original magnitude but is forced to precess in the transverse plane about $\boldsymbol{B}_0$, inducing a signal in the receiver coil, the so-called *free-induction decay* (FID). The transverse magnetization (also called coherence) is the sum of many contributions, originating from different spin positions inside a molecule and from different regions in space. Both effects — which were separated into microscopic and macroscopic above — generate a certain spread of local fields $\boldsymbol{B}$, and of Larmor frequencies ω, so that the individual contributions of the initial (maximum) magnetization begin to "fan out" (note that the plot in Figure 1 must be understood as the *rotating frame*, i.e., one assumes a point of view that is rotating with the carrier frequency of the transmitting radio field about the z axis. In the rotating frame, some partial magnetizations (also called *isochromats*) precess faster, while others precess slower than the average. Eventually, their vectors add up to zero and no signal is detected in the receiver. The 180° pulse "flips" all isochromats around, so that the faster ones suddenly find themselves at the trailing edge, but because the physical reason for their precession frequency has not changed, they begin to catch up with the slower ones. This process goes completely unnoticed by the operator of the NMR spectrometer, because the isochromats remain in a decoherent state relative to each other, and their vectorial sum is still zero. Only after a time which is exactly twice the interval between the two rf pulses do they get in phase again, and in this moment, the broad arrow, the complete magnetization, is regained, but now in the direction opposite to where it had pointed to immediately after the first rf pulse. This is because the 180° rotation was performed about the x axis, as depicted in Figure 1. By changing the *phase* of the rf pulses, i.e., by changing the axis about which the rotation of the magnetization vector takes place, as discussed above, the final direction of this arrow can be chosen at will. The recovered magnetization is the so-called *spin echo*. The second half of this echo is identical to the FID at the beginning of the sequence, apart from the actual signal amplitude, which has decreased due to microscopic or macroscopic processes that have happened in the system during the time between the FID and echo acquisition. Comparing echo and FID amplitudes can give a first starting point for analyzing dynamics that are taking place in the sample; using the echo and continuing the procedure by applying more rf pulses allows one to recycle the very same magnetization many times and extract even more information from its behavior. Imaging is achieved by appropriate manipulation of the magnetization in the intervals between the rf pulses.

2.2. k Space

The encoding of positional information into the NMR signal, and the various experimental schemes that allow this encoding in an optimized way to match the needs of the desired study, are frequently discussed in terms of *k space*. This concept has been borrowed from scattering experiments (light, X-ray, neutron scattering) where the structure of matter is revealed by interactions with the wave nature of incoming radiation, a function of their *wavelength* or, in accordance with the structure of the scattering pattern, its reciprocal, the *wave vector* or $\boldsymbol{k}$ *vector*. Note that in NMR imaging, it is not the wave vector of the rf irradiation itself, but rather a spatially periodic dependence of the magnetization that is generated by a suitable manipulation of the spin system in the presence of a spatially dependent magnetic field. The actual image information, i.e., the distribution of magnetization in real space $\boldsymbol{r}$, can be obtained from a Fourier transformation of the signal in dependence on the reciprocal vector $\boldsymbol{k}$. Because of this fundamental relationship, it appears worthwhile to gain some elementary understanding about magnetization encoding in k space.

The spatial dependence of the NMR signal is introduced straightforwardly by allowing the magnetic field strength in Equation 6 to become dependent on position $\boldsymbol{r}$:

$$\omega(\boldsymbol{r}) = \gamma\,|\boldsymbol{B}(\boldsymbol{r})| = \gamma(|\boldsymbol{B}_0| + \boldsymbol{g}\cdot\boldsymbol{r}) \tag{7}$$

Here, the magnetic field strength has been split into a homogeneous part $\boldsymbol{B}_0$ and a spatially dependent part, expressed by the first derivative of the magnetic field strength, the gradient vector $\boldsymbol{g}$, which renders the resonance frequency proportional to the position of the spin $\boldsymbol{r}$. This expression resembles the conditions of a typical MRI setup, where the additional, spatially varying field $\boldsymbol{g}\,\boldsymbol{r}$ is only a small correction to $\boldsymbol{B}_0$. Strictly speaking, the gradient has the properties of a 3×3 tensor because the derivatives of all three orthogonal components of $\boldsymbol{B}$ must be computed. However, $\boldsymbol{B}$ is usually taken as pointing in the z direction, and its x/y components are negligible just like their spatial derivatives. Note that the product $\boldsymbol{g}\,\boldsymbol{r}$ indicates that only those components of the position vector which are parallel to the gradient direction possess a different resonance frequency.

Equation 7 expresses the Larmor frequency for a subset of spins at position $\boldsymbol{r}$. In an actual experiment with the purpose of generating an image of an extended object, a large number of these subsets, or *isochromats*, will give rise to a superposition of frequencies $\omega(\boldsymbol{r})$, weighted by the distribution of spins in the object. This distribution can formally be written by the density function of position: $\varrho(\boldsymbol{r})$. Since most NMR experiments consist of an initial rf pulse followed by one or more encoding periods with different properties, the signal detected at the end of such a sequence that includes a magnetic field gradient will contain a superposition of the spatial information that cannot be directly translated into an image. In order to generate the image of the object, one must modify the weighting function (the effective gradient $\boldsymbol{g}$ turned on during the time t, hence $\boldsymbol{g}t$) and record the signal for each weighting function separately, before reconstructing the actual spin density distribution. This approach corresponds to the scattering experiment with an incoming wave of vector $\boldsymbol{k}$. However, there is no direct NMR equivalent for a two-dimensional detector that allows, for instance, an X-ray diffraction pattern to be recorded in one shot. The values of the $\boldsymbol{k}$ vector rather must be set one by one and the NMR response signal acquired separately. [Note that Eq. 7 implies that there is a direct proportionality between the spin density distribution $\varrho(\boldsymbol{r})$ and the frequency distribution $\varrho[\omega(\boldsymbol{r})]$, which allows reconstruction of the projection of the spin density parallel to the direction of the gradient $\boldsymbol{g}$ if the latter is applied during data acquisition; this particular method is discussed in Section 2.3]

On the other hand, there is one major difference which makes NMR a particularly strong tool for investigating matter: unlike the classical scattering techniques, which only provide an intensity measure as their result, the NMR information is a complex number and also contains the phase of the signal (proper synchronization and detection of the induced voltage in quadrature allow the determination of the signal phase at any instant during acquisition). By applying field gradient pulses, the position can be encoded into the phase of the signal as a function of $\boldsymbol{k}$, while its magnitude and frequency remain available for assigning other information (see Sections 2.4, and 3.3).

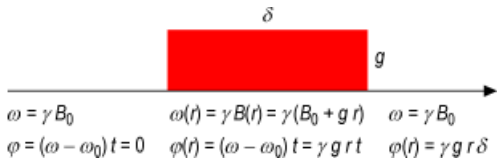

Figure 2. Effect of a pulsed magnetic field gradient of strength $\boldsymbol{g}$ on the phase of a signal contribution originating from spins at position $\boldsymbol{r}$. Prior to the gradient pulse, all spins experience the constant magnetic field B0 and thus have the same Larmor frequency; the phase shift relative to this average value is zero. During the gradient pulse, the field becomes position-dependent and a phase shift ϕ is accumulated that is proportional to position $\boldsymbol{r}$ and time t. After the gradient pulse (i.e., after completion of the phase-encoding step), the spins memorize their individual phase shift? $\phi(\boldsymbol{r}) = \gamma\,\boldsymbol{g}\boldsymbol{r}\delta$.

Figure 2 demonstrates the effect of a linear field dependence that is introduced in the form of a constant gradient $\boldsymbol{g}$ which is applied only during a certain interval δ. This is called a — hence the frequently used term "pulsed field gradient (PFG) NMR" — and it can be of arbitrary shape, but a rectangle with sharp edges is often used and facilitates the discussion. The Larmor frequency is then simply switched between its value in the homogeneous field, $\boldsymbol{B} = \boldsymbol{B}_0$, $\omega_0 = \gamma\boldsymbol{B}_0$, before and after the gradient pulse, and the position-dependent value $\omega(\boldsymbol{r}) = \gamma\,|\boldsymbol{B}(\boldsymbol{r})| = \gamma(|\boldsymbol{B}_0| + \boldsymbol{g}\cdot\boldsymbol{r})$ during the gradient pulse. It is convenient to subtract the reference value, given

by the "average" field $\omega_0 = \gamma \boldsymbol{B}_0$ so that one obtains the frequency difference relative to an (arbitrarily chosen) position $\boldsymbol{r} = 0$:

$$\omega(\boldsymbol{r}) - \omega_0 = \gamma \boldsymbol{g} \cdot \boldsymbol{r}. \tag{8}$$

In NMR, data acquisition is demodulated with the carrier frequency corresponding to the magnetic field strength in use ω_{rf}, which is chosen usually close to ω_0, so that the frequency difference of Equation 8 can even be detected directly (see Section 2.2).

The bottom line in Figure 2 represents the phase ϕ of the signal, which in the rotating frame of reference is defined as zero in a homogeneous magnetic field. As soon as the gradient $\boldsymbol{g}$ is switched on, the phase angle grows linearly with time, according to $\phi(\boldsymbol{r}) = (\omega(\boldsymbol{r}) - \omega_0)t$. Because of the dependence of the Larmor frequency on position $\boldsymbol{r}$, this phase angle also depends on the position. After the gradient pulse is switched off, the difference in frequencies has led to an accumulated phase shift $\phi(\boldsymbol{r}) = (\omega(\boldsymbol{r}) - \omega_0)\delta$ with respect to the reference, and this will not change as long as the magnetic field remains homogeneous. Thus, the resulting phase shift of spins at position $\boldsymbol{r}$ following a gradient pulse of duration δ and strength $\boldsymbol{g}$ can be written as

$$\phi(\boldsymbol{r}) = \gamma\, \boldsymbol{g} \cdot \boldsymbol{r}\, \delta. \tag{9}$$

Since the signal generated by the precessing magnetization is a complex number, the phase shift created by this gradient pulse leads to an NMR signal which is the multiplication of the unperturbed NMR signal with the phase factor $\exp[i\phi(\boldsymbol{r})]$. To compute the total signal arising from all spins within the sample with a spin density distribution $\varrho(\boldsymbol{r})$, the different phase factors must be integrated over the entire volume:

$$\begin{aligned} S(\boldsymbol{k}) &= \int \varrho(\boldsymbol{r}) \exp[i\phi(\boldsymbol{r})]\, \mathrm{d}\boldsymbol{r} \\ &= \int \varrho(\boldsymbol{r}) \exp[i\gamma \boldsymbol{g} \cdot \boldsymbol{r} \delta]\mathrm{d}\boldsymbol{r} \\ &= \int \varrho(\boldsymbol{r}) \exp[i2\pi \boldsymbol{k} \cdot \boldsymbol{r}]\mathrm{d}\boldsymbol{r}. \end{aligned} \tag{10}$$

In this equation, the substitution $k = \frac{1}{2\pi}\gamma g \delta$ has been made to introduce the Fourier conjugate variable to $\boldsymbol{r}$. This is because Equation 10 formally is a Fourier transformation. The shape of the sample $\varrho(\boldsymbol{r})$ is the relation that is sought-after, and it can be derived by applying the inverse Fourier transformation to the signal function $S(\boldsymbol{k})$:

$$\varrho(\boldsymbol{r}) = \int S(\boldsymbol{k}) \exp[-i2\pi \boldsymbol{k} \cdot \boldsymbol{r}]\mathrm{d}\boldsymbol{k} \tag{11}$$

2.3. Phase Encoding and Frequency Encoding

Equation 11 constitutes the fundamental relationship for NMR imaging: in order to find the one-dimensional spin density $\varrho(\boldsymbol{r})$, one must determine the signal as a function of the wave vector $S(\boldsymbol{k})$.

For a proper application of the inverse Fourier transformation, $S(\boldsymbol{k})$ needs to be known as a continuous function. In practice, it is the fast Fourier transform (FFT) which is performed rather than the full, analytical Fourier transform, so that the sampling of $\boldsymbol{k}$ space at discrete, equidistant steps (typically 32, 64, 128) is performed.

The recipe for the first building block of NMR imaging, *phase encoding*, thus looks like this: apply a phase gradient of effective area $\boldsymbol{k}$, acquire the signal $S(\boldsymbol{k})$, repeat for a number of different equidistant values of $\boldsymbol{k}$, and perform the inverse Fourier transformation to reconstruct the spin density function of the sample $\varrho(\boldsymbol{r})$. The variation of gradients is symbolized by a diagonal line in Figure 3. In principle, the pulsed gradient can be applied at arbitrary times between the rf pulses.

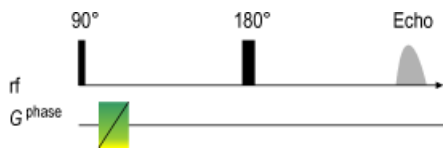

Figure 3. Phase-encoding scheme in one dimension. A magnetic field gradient pulse is applied and is varied between subsequent scans (symbolized by the diagonal line). The actual timing of the gradient is arbitrary as long as it is placed between the rf pulses and before the acquisition of the echo signal.

Note that variation of the phase encoding gradient along one direction, say k_x, allows the reconstruction of the profile only in this direction, $\varrho(x)$. An example of this is shown in Figure 4, where the projection of a cylindrical object (such as a test tube filled with water) is depicted with the aid of simulated data. A series of FIDs is drawn in succession, a usual way of saving the data in a single, long vector file in which each FID is being acquired in the presence of a particular gradient value, so that the range $-k_{x,\max} \ldots k_{x,\max}$ is covered and $k_x = 0$ is in the center,

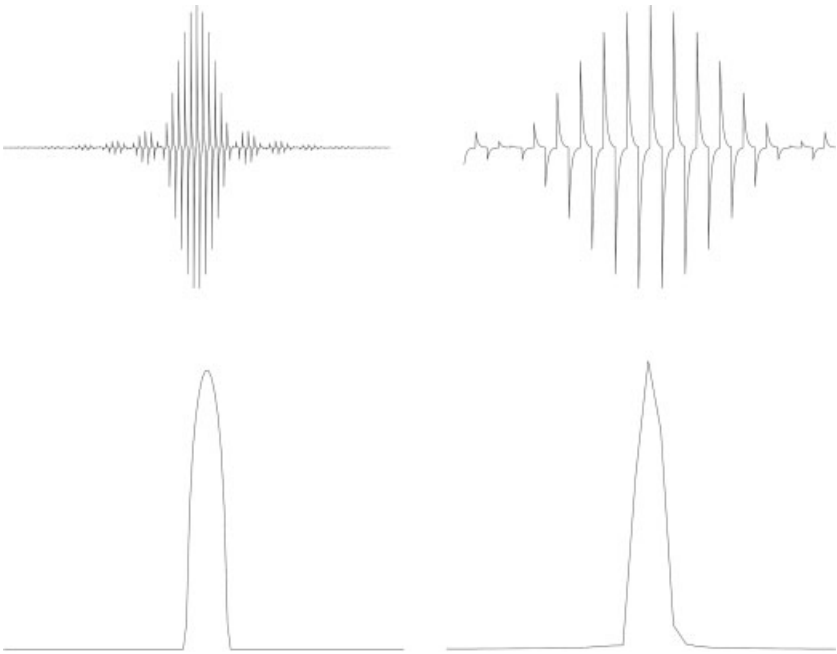

Figure 4. Simulated NMR data from a cylindrical object projected onto a direction perpendicular to the cylinder axis. Left: 128 steps of k_x were applied and the succession of FIDs plotted; $\boldsymbol{k}$ space is covered sufficiently well, i.e., the signal has decayed to zero at the largest values of k_x. The Fourier transform of the series of FIDs (taken from the first points or from the integrals over each FID) is a satisfactory representation of the projection of the sample shape. Right: only the central 32 k_x values were acquired; the Fourier transform suffers from bad resolution, but also from a nonvanishing baseline because the signal at the largest k_x values had not vanished.

giving maximum signal intensity. Figure 4 also demonstrates the effect of an insufficient coverage of $\boldsymbol{k}$ space on the image quality: if at the largest used value of k_x, the signal has not yet vanished, cutoff artefacts arise in the image (see Section 2.6).

The foremost advantage of phase encoding is the fact that the signal is acquired in the presence of a homogeneous magnetic field; the signal thus contains the full spectroscopic information, and this can be exploited to generate an image with individual spectra contained in each voxel representing the local chemical environment. This is one of the most powerful assets of NMR imaging. In one important example, an inhomogeneous distribution of a mixture of liquids can be imaged while obtaining the full spectroscopic dimension to distinguish the different phases and/or chemical species in the mixture. By this means, chemically selective, spatially resolved monitoring can be employed to study complex mixtures. The disadvantage of phase encoding is its inherent slowness: the signal function at each discrete value of $\boldsymbol{k}$ must be acquired separately, and a sufficient waiting time introduced to allow the magnetization to relax back to equilibrium (see Chap. 4 for strategies to reduce this waiting period).

One way out of this dilemma is the so-called *frequency encoding* technique. It is often referred to as an approach which is quite different from phase encoding, but in fact it is only the timing of $\boldsymbol{k}$ space sampling which is modified: in frequency encoding, the signal is acquired while the field gradient is present (see Fig. 5); it is then referred to as a *read gradient*. Equation 9 is still valid, so that each subsequent point of acquisition at a given time t is subject to the effect of a longer read-gradient field and thus contains a different distribution of phase angles. The full acquisition of one line in $\boldsymbol{k}$ space takes only a few milliseconds; Fourier transformation of this signal $S(\boldsymbol{k})$ again yields the profile of the object in the corresponding direction $\varrho(\boldsymbol{r})$, so that Equation 11 also maintains its validity. Since in frequency encoding the time t is varied at constant gradient strength, it is also possible to write the signal of the frequency encoding technique

as $S(t)$, and its Fourier transformation directly renders the frequency distribution of spins in the sample. Because of $\omega = \gamma \boldsymbol{gr}$, this frequency distribution is proportional to the spin density distribution of the object. The difference in the *phase encoding* method is that the signal is acquired after it was exposed to a gradient field of duration δ and strength $\boldsymbol{g}$. Usually, the gradient strength $\boldsymbol{g}$ is varied, and for each such step a new signal is utilized, formed either by excitation or echo formation.

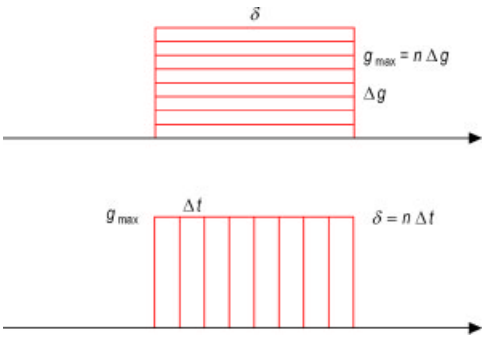

Figure 5. Demonstration of the formal equivalence of phase encoding and frequency encoding. Phase encoding (top): the gradient of duration δ and strength $g_{\max}$ is split into eight equal steps Δg. The signal is acquired n times (here: $n = 8$), following a gradient of strength $i\Delta g$ where i is varied from 1 to n. Frequency encoding (bottom): the gradient strength $g_{\max}$ is maintained but the total duration δ is split into eight equal steps Δt. The signal is acquired n times (here: $n = 8$), following a gradient of duration $i\Delta t$ where i is varied from 1 to n. In both cases, the signal corresponding to the same $n = 8k$ values between zero and $k_{\max}$ is sampled, where $k_{\max} = (2\pi)^{-1}\gamma g_{\max}\delta$.

While frequency encoding saves time because it samples one dimension in $\boldsymbol{k}$ space in a single acquisition, its disadvantage is that the spectral information of the spin system is masked by the spatial encoding; the result is a superposition of chemical shift and positional information. In other words, an object filled with a substance that has a spectrum consisting of two lines will render a double image because the physical reason for the frequency shift cannot be separated straightforwardly. However, often substances with simple spectra such as water are considered, or the separation between different spectral lines is negligibly small. Under such conditions, frequency encoding is often the method of choice.

2.4. Slice Selection

In many real studies, a cross-sectional image of an object contains all the necessary information that the researcher is interested in. Sometimes it might even be desirable to generate an image of only a partial volume of the object, or to determine a spectrum at a particular position. However, the methods discussed so far affect all spins within the volume that is common to the gradient coil unit and the resonator coil, and do not focus on a partial volume. Volume selection, or more generally *slice selection*, is indeed possible; again, it exploits the fundamental relation $\omega = \gamma \boldsymbol{B}$, but in a slightly different sense. Up to now, it has been implicitly assumed that all rf pulses in the imaging sequences affect all spins in the same way. This is strictly true only for a *hard pulse*, i.e., a pulse with a bandwidth well exceeding the range of Larmor frequencies in the sample. The bandwidth of an rf pulse is approximately given by the reciprocal of its duration: a short pulse has a broad frequency spectrum. More precisely, the frequency spectrum of an rf excitation is the Fourier transform of its shape function. But because of $\omega(r) - \omega_0 = \gamma \boldsymbol{g} \cdot \boldsymbol{r}$, frequency is proportional to position if a magnetic field gradient $\boldsymbol{g}$ is present. In principle, a slice of arbitrary shape can be excited if the pulse is modulated by a function which is the inverse Fourier transformation of the shape. A volume is likewise excited by combining rf pulses and orthogonal gradients, and even arbitrary shapes can be selected [16]. For the most appropriate excitation profile, a rectangular slice, this is particularly simple — the Fourier transform of a rectangle is the sinc function, $\sin x/x$ (see Fig. 6). Generating an rf pulse of the proper length which is modulated by a sinc function will thus excite a rectangular slice of the sample and leave the remaining spins unaffected. In practice, the function is cut at a certain number of periods to limit its duration, and this results in a somewhat imperfect shape of the slice, which is usually tolerated (see Fig. 6). In typical imaging, the *shaped pulse* or *soft pulse* has a duration on the order of milliseconds, compared to the *hard pulse* of some tens of microseconds. The terms "hard" and "soft" can thus also be understood in the sense of the power level of the pulses, which necessarily is much higher if a very short

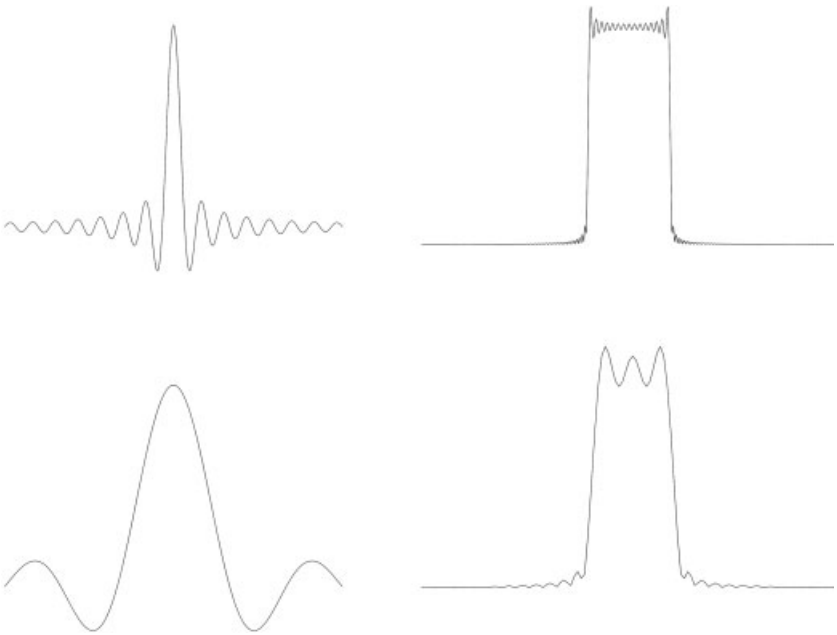

Figure 6. Soft rf pulses (left) in the shape of a sinc (sin x/x) function, and their Fourier transforms (right), equivalent to the excited slice in the presence of a constant magnetic field gradient. The well-defined sinc function (top) produces a slice excitation that deviates only slightly from a perfect rectangle. A cutoff sinc function, such as that shown in the bottom part of the figure, produces a distorted slice profile which is, however, often sufficient for slice selection in multidimensional imaging.

pulse has to achieve the same 90° rotation of the magnetization.

The shape of any rf pulse can be chosen in such a way that the excitation profile is a rectangular slice. In the light of experimental restrictions, which often ask for pulses as short as possible, the slice shape will never be perfect and also depends on the pulse's flip angle. For instance, the commonly used 90° pulse is still acceptable, while a 180° pulse produces a good profile only if it is used as a refocusing pulse, but not when it is supposed to act on the longitudinal component of the magnetization M_z. Sometimes, pulses of even smaller flip angles are used that provide a better slice selection (for a discussion of imaging with small flip angles, see Section 4.2.

2.5. 2D and 3D Imaging

Obtaining two- or three-dimensional NMR images is achieved by employing different combinations of the aforementioned techniques. For instance, pure phase encoding is extended straightforwardly from the simple module shown in Figure 3, the only requirement being that all gradient pulses are stepped independently of each other, and are applied either simultaneously or in series. Equation 11 already contains this possibility since $\boldsymbol{r}$ and $\boldsymbol{k}$ are written as vectors; however, in Equation 12 the three orthogonal components are split and — for the sake of clarity — the same relation is expressed by a threefold integral to highlight the three-dimensional nature of the experiment:

$$\varrho(x,y,z) = \int\int\int S(k_x,k_y,k_z)\exp[-i2\pi k_x x]\exp[-i2\pi k_y y] \cdot\exp[-i2\pi k_z z]\mathrm{d}k_x\mathrm{d}k_y\mathrm{d}k_z \quad (12)$$

The disadvantage of long experimental times in phase encoding becomes almost prohibitive for 3D experiments: to generate a three-dimensional image with a satisfactory resolution of, say, 64 × 64 × 64 points, an equivalent number of 64^3 individual signals must be acquired. Assuming a repetition time of 1 s between acquisitions, this corresponds to an experimental time of three days.

Equation 12, however, allows one of the three dimensions to be scanned by frequency encod-

ing. This has turned a combination of frequency encoding in one dimension and phase encoding in one or two further dimensions into one of the most widespread imaging techniques. Figure 7 gives a schematic example of a suitable pulse sequence, where the succession of gradients is arbitrary as long as all phase encoding steps are completed before the beginning of signal acquisition.

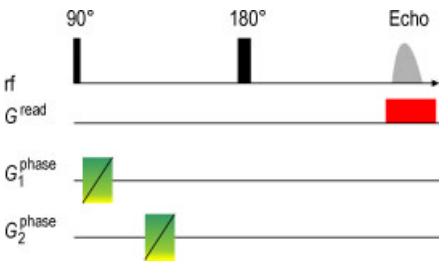

Figure 7. Schematic plot of a basic three-dimensional imaging pulse sequence with frequency encoding along one axis (read), and phase encoding along the two remaining orthogonal directions. The choice of directions is arbitrary, as is the position of the phase gradients within the sequence.

Reconstruction of a two- or even three-dimensional spin density distribution by employing only frequency encoding is possible by varying the direction of the gradient by rotation of the sample or the gradient coil, or by rotating the gradient fields by appropriate combinations of g_x, g_y, and g_z gradients. The technique is known as *backprojection/reconstruction* and is similar to image processing routines used for generating 3D X-ray images in computer tomography (CT). While the advantage is that a crude picture of the sample can be obtained right away long before image acquisition is completed, some disadvantages are the lower signal-to-noise ratio, resolution, and image quality, and that it does not allow the assignment of spectral features in the object. The reason for these compromises is related to the fact that the image function $\varrho(\boldsymbol{r})$ is extracted by an algorithm different from the inverse Fourier transformation.

By using all three building blocks of phase encoding, frequency encoding, and slice selection, a very common NMR imaging sequence can be constructed which is routinely employed for acquiring slices of an extended object (see Fig. 8). It consists of a slice-selection gradient in one direction, a frequency gradient in a second, and a phase-encoding gradient in the third direction. The directions are orthogonal to each other and their order is arbitrary, hence the usual labeling with "read, phase, slice" instead of "x, y, z".

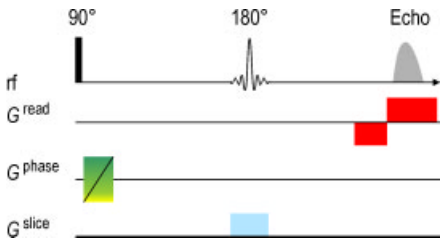

Figure 8. Typical basic three-dimensional imaging sequence with slice selection, frequency encoding, and phase encoding in three orthogonal directions. The compensating lobe for the read gradient is drawn as a rectangle with negative intensity directly before the actual read gradient. The shape of the 180° rf pulse is drawn schematically to indicate that a soft pulse is used.

A final but important comment must be added to the discussion of the sequence in Figure 8. The read gradient is preceded by a negative lobe of half its area. The reason for this is that one wants to scan the whole $\boldsymbol{k}$ space dimension with the read gradient. If one switched on the gradient and began acquisition immediately, the first acquired point would represent $\boldsymbol{k} = 0$ and only positive values of $\boldsymbol{k}$ would be scanned. By applying a negative lobe before acquisition, the origin is moved toward a negative value before acquisition is begun; applying a lobe with half of the negative area of the read gradient means that the condition $\boldsymbol{k} = 0$ appears at the center of the acquisition window. The timing of the sequence is such that also the echo generated by the rf pulses appears in this instance. It is called *spin echo* and refocuses the effect of constant background gradients, but also the spread in ω brought about by the properties of the sample spectrum (see also discussion of Fig. 1). The combination of negative and positive lobes of g_{read} also produces an echo, the *gradient echo*. It appears at the point in time when $\boldsymbol{k} = 0$, or, in other words, when $\int g_{\mathrm{read}} \mathrm{d}t = 0$, i.e., the total effect of all gradients in the read direction vanishes. The integral is the mathematical definition of the area of the gradient, or the zeroth moment of the gradient $\boldsymbol{m}_0$. The general nomenclature will be put to use in Section 5.2.

The equivalence of phase encoding and frequency encoding is best seen in relating k space and r space to each other, which is visualized for the two-dimensional case in Figure 9. The only essential feature is the equidistant scanning of points in k space, indicated by a rectangular grid, which generates an image in r space of likewise equidistant points. The separation between points needs not be the same in both directions, the order of acquisition is – at least in principle – arbitrary, and each point can be part of a phase- or frequency-encoding scheme.

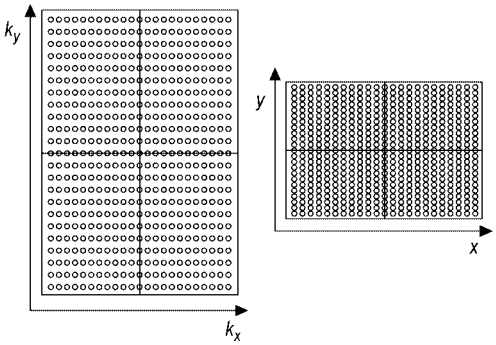

Figure 9. Scanning of 2D k space and the equivalent reconstruction in 2D r (position) space. In this example, the same number of points is sampled in k_x and k_y directions, but the total range of k_y values is 50 % wider; as a consequence, the image resolution along y is 50 % better than along x, while the imaging window along y (field of view, FOV) is reduced by the same amount.

2.6. Spatial Resolution

To demonstrate how the Fourier relation works, Figure 10 shows how a two-dimensional object — a letter "i" in a circle — would look like in an NMR image, that is, its spin density function $\varrho(x,y)$ is shown, and the original signal function which is obtained by varying $\boldsymbol{k}$, $S(k_x, k_y)$. While $S(k_x, k_y)$ is a complex function, only its real part is plotted. Because of the reciprocal relationship between $\boldsymbol{r}$ and $\boldsymbol{k}$, one can roughly say that the signal of high intensity in the vicinity of $\boldsymbol{k} = 0$ contains information about the general shape of the sample, and the small contributions at large $\boldsymbol{k}$ values are responsible for the fine structure, i.e., the resolution of the image. Although this is an oversimplified view, it is often helpful to keep this inverse relationship in mind and to estimate which effect the distortion of a particular part of the $\boldsymbol{k}$ space data will have on the $\boldsymbol{r}$ space image. In principle, only an analytical Fourier transformation of a continuous signal function $S(\boldsymbol{k})$ renders a perfect result; the compromise of a finite sampling of $\boldsymbol{k}$ space gives rise to imperfections in the NMR image, which arise in addition to potential technical shortcomings such as hardware limitations or noise.

The choice of gradient strength and duration can be optimized if the size of the object is known. From Equations 9 and 10 it becomes clear that too large a gradient will lead to an ambiguous assignment of positions if the maximum phase angle exceeds 2π. One must therefore make sure that the difference between phase shifts generated for spins at the extreme ends of the sample is smaller than 2π for each different value of the gradient. In other words, $2\pi > 2(\gamma g_{\max}\delta r_{\max})/(n-1)$ if the gradient is varied in the range $-g_{\max} \ldots + g_{\max}$ in n steps. The coverage in $\boldsymbol{r}$ space achieved with a particular set of gradient values is called the field of view (FOV) and is obtained from

$$\mathrm{FOV} = \frac{n-1}{2}\,\frac{2\pi}{\gamma g_{\max}\delta} \tag{13}$$

For n gradient steps (n different k values), the image function $\varrho(r)$ also possesses a resolution of n points as a consequence of the FFT procedure. Naturally, a larger $g_{\max}$ results in a better resolution if the number of steps n is chosen in such a way that the FOV slightly exceeds the object size itself (see Fig. 9). Note that choosing too small a field of view will lead to so-called aliasing, i.e., those picture elements that correspond to a phase shift that exceeds the range $\pm\pi$, will appear at the other side of the image, so that superposition of two or more signal contributions into the image (backfolding) can occur. Note that Equation 13 is again valid for phase encoding and frequency encoding alike; in the latter case, n is the number of time points during acquisition in the presence of the magnetic field gradient.

A much more detailed introduction to the basics of NMR imaging can be found in a number of textbooks and monographs [17 – 20]. These also cover most of the aspects presented throughout the following chapters. While [21, 22] are dedicated to the multitude of pulse sequences used for MRI, [23, 24] provide a collection of

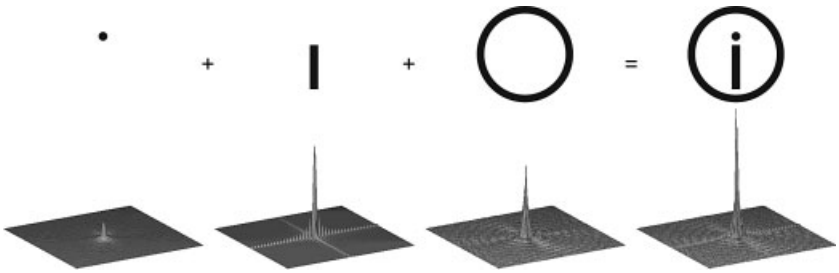

Figure 10. Two-dimensional objects in position space (top row) and their Fourier transform, corresponding to the shape of the signal function $S(k_x,k_y)$ (bottom row). The actual object, a letter "i" inside a circle, is shown as the sum of three individual objects. Likewise, the Fourier transform is additive and the signal functions corresponding to each of the objects are shown for comparison.

applications from different, mostly nonmedical fields of research.

3. Image Contrast

3.1. Relaxation

At first sight, the image brightness appears to be a direct measure of NMR signal intensity obtained under the particular pulse sequence used in image acquisition, which intuitively is considered to be the number density of spins, e.g., of protons in the sample. However, this brightness does not only reflect spin density, but is weighted by the chemical constitution, hardness or softness of the material, diffusion coefficient, translational motion, or molecular motion on a much faster timescale, which actually defines the contrast of an image. The contrast highlights the heterogeneity of the sample with respect to the chemical and physical properties mentioned. This is because these properties reflect on NMR parameters such as NMR signal amplitude, frequency, T_1 and T_2 relaxation times, and line shape. One can learn much more about the properties of a sample since the NMR imaging sequence can be designed accordingly to obtain the desired contrast. The wide range of contrast that can be imposed on a spin-density image is the unique strength of NMR imaging. On the other hand, quantitative NMR imaging, i.e., the determination of local spin densities and therefore substance concentrations, is not straightforward because of these manifold influences, and some care must be taken to derive the spin density distribution in an object.

The most obvious influence on image contrast is given by the relaxation parameters, in particular the spin – lattice relaxation time T_1 and the spin – spin relaxation time T_2. Both describe the return of the magnetization to its equilibrium value which is determined by the strength and direction of the external magnetic field. Considering that any encoding of information requires the manipulation of the magnetization that removes it from its equilibrium, one can conclude that full relaxation removes the memory about any previous encoding events and is the requirement for performing a new experiment. Imaging sequences such as those drawn in Figures 3, 7, and 8 can then be repeated only with sufficient waiting times in between. Incomplete relaxation, on the other hand, implies that the spin system potentially carries information from previous encoding steps. This can either be a nuisance or an advantage, as will be demonstrated in Chapter 4. In any case, incomplete relaxation reduces the signal intensity and therefore the image brightness compared to the ideal, fully relaxed state.

Historically, relaxation effects were introduced phenomenologically by amending the Bloch equation, i.e., the equation describing the time dependence of the magnetization vector (Eq. 1), by a decaying term:

$$\frac{\mathrm{d}\boldsymbol{M}}{\mathrm{d}t} = \gamma \boldsymbol{M} \times \boldsymbol{B} - \mathbf{R}(\boldsymbol{M} - \boldsymbol{M}_0) \tag{14}$$

Here, $\mathbf{R}$ denotes the relaxation matrix that has the diagonal elements $R_{xx} = R_{yy} = 1/T_2$, $R_{zz} = 1/T_1$. Putting Equation 14 in words, the transverse magnetization M_x will decay from its initial value to zero with a time constant T_2, while at the same time the longitudinal component M_z returns to its equilibrium value M_0, with a time constant T_1. The definition of these time con-

stants is based on the assumption that the relaxation process follows an exponential behavior. This is often the case, but not always; if a different time dependence is found, the relaxation time loses its meaning in the stricter sense. It is then formally replaced by a function that can be split into a sum of many exponentially decaying functions, or a different measure is used; for practical reasons, the time during which the magnetization decays to e^{-1} of its initial value is often simply called the relaxation time T_1 or T_2.

The theory of relaxation is covered in most standard introductory books, e.g., [1 – 6], and is not repeated here. In principle, relaxation processes are determined by the interaction between spins. This is usually the dipolar coupling in the case of protons, but other types of interactions can also contribute significantly. Once the spin system is brought out of equilibrium by an rf pulse, it begins to return to the initial state, i.e., the populations of the energy levels will return to the Boltzmann distribution of Equation 5. This is only possible if energy is exchanged between neighboring spins, and an energy exchange generally requires a change of orientations, and therefore motion. Transverse relaxation happens simultaneously, but is somewhat different since it is described by a loss of coherence of magnetization, which does not necessarily require a change of the total energy. Because of the additional mechanisms leading to this coherence loss, T_2 is often shorter than T_1. The T_2^* relaxation time is *not* an intrinsic property of the material itself and is called the effective transverse relaxation time. Inhomogeneity in static magnetic fields and the sample's magnetization due to susceptibility variations are the cause for T_2^* and the resulting coherence loss is reversible, while the true T_2 describes an irreversible decay process. T_2^* is generally shorter than T_2 and often becomes important in imaging experiments where a "perfect" magnetic field cannot be achieved. It is then the time constant that describes the free induction decay, but the lost coherence can be recovered at the center of a spin echo, as is drawn schematically in Figure 1. The effect of relaxation can be demonstrated with the help of Figure 11 and Equations 10 and 12.

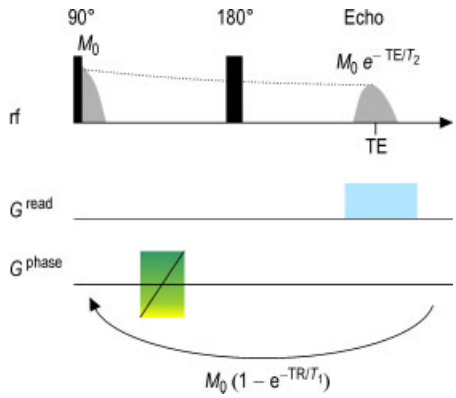

Figure 11. Schematic plot of a basic two-dimensional imaging sequence with relaxation effects. During the encoding period, the magnetization remains in the transverse (xy) plane and decays by a factor $\exp(-\mathrm{TE}/\mathrm{T}_2)$, where TE is the echo time. After signal acquisition, which occurs at the time of the echo, the sequence is repeated following a waiting time TR (repetition time), during which the longitudinal magnetization approaches its equilibrium exponentially according to $[1-\exp(-\mathrm{TR}/\mathrm{T}_1)]$.

It is sufficient to remember that the components of the magnetization in the transverse plane, M_x and M_y, experience T_2 and T_2^* relaxation, while the component parallel to the magnetic field M_z approaches its equilibrium with T_1. The correct expression for the signal intensity can then be rewritten as:

$$S\,(k_x,k_y,k_z) = \int\!\!\int\!\!\int \varrho\,(x,y,z)\ \exp\left[-\mathrm{TE}/T_2\,(x,y,z)\right] \left(1-\exp\left[-\mathrm{TR}/T_1\,(x,y,z)\right]\right) \exp\left[i2\pi k_x x\right]\ \exp\left[i2\pi k_y y\right] \exp\left[i2\pi k_z z\right] \mathrm{d}k_x\,\mathrm{d}k_y\,\mathrm{d}k_z \tag{15}$$

This equation assumes the general case where the parameters T_2 and T_1 depend on position, which is often the case, in particular if the material composition in an object is heterogeneous. Under these circumstances, the actual spin density $\varrho(x,y,z)$ will be damped by a position-dependent factor. Without knowledge of the dependence of $T_1(x,y,z)$ and $T_2(x,y,z)$, which is generally not given, the spin density cannot be retrieved unless the conditions $\mathrm{TE} \ll T_2$ and $\mathrm{TR} \gg T_1$ are met for the whole range of relaxation times in the sample. These two time intervals are called "echo time" (TE) and "repetition time" (TR) where the nomenclature is borrowed from medical imaging.

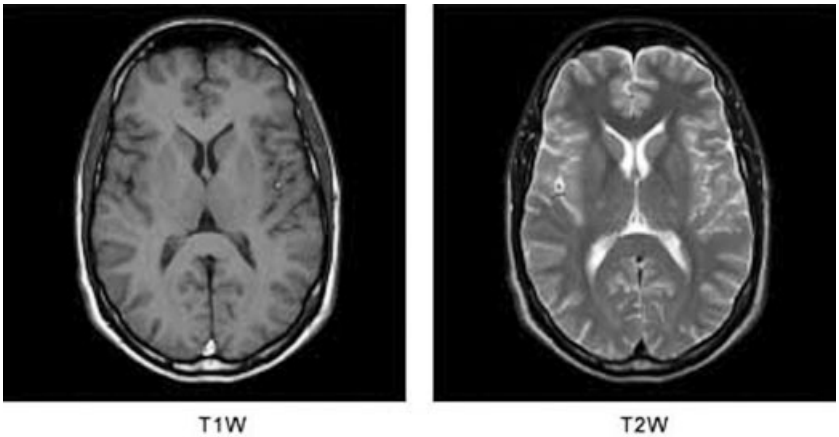

Figure 12. . Images of a human brain with T_1 weighting (left) and T_2 weighting (right). T1 weighting is brought about by using a TR that is short relative to the average T_1 of the tissue. Gray matter has a slightly longer T_1 relaxation time than white matter, making it appear darker in the image; the contrast depends also on the field strength used. T_2 weighting is achieved by employing a long TE. Water or, more generally, free liquid has a long T_2 and appears bright. The major potential of medical relaxation weighted imaging is in the detection of lesions and cancerous tissue, which contains more water and therefore usually has longer relaxation times. (Reproduced from http://www.thebraincentre.org/neuroradiology/mri/image/mri03.jpg, The Brain Center, Canossa Hospital, Hong Kong.)

An example where different repetition and echo times lead to different contrast is shown in Figure 12.

One of the first criteria for determining the feasibility of a particular imaging experiment is a good estimate for the relaxation times of the system under study, which then must be compared with the times required for executing the pulse sequence, frequently determined by hardware limitations (see Chap. 6). In bulk water and most organic liquids, one usually has $T_1 = T_2$ and both are in the range of several seconds. This is sufficiently long for executing all types of imaging pulse sequences, but might have the disadvantage that repetition times become prohibitively long, or alternatively the image is always T_1-weighted. The addition of soluble salts that form a hydration shell can help, because the extra relaxation mechanism introduced by the ions can reduce the relaxation times by orders of magnitude [25, 26]. Such salts are employed as contrast agents in routine medical MRI, and their design towards an increased specificity currently represents a dynamic research area [27].

However, for liquids with a high concentration of ions, or very viscous liquids, the opposite problem can arise if the relaxation times become too short for completing a full pulse sequence and to acquire the remaining signal. The extreme case is represented by solids, which have a very long longitudinal relaxation time T_1 (anything from seconds to hours, depending on the rigidity and order of the molecular structure), while the transverse relaxation time T_2 becomes very short, often being only on the order of 10 µs. To a lesser degree, this also happens if liquids are in contact with large specific interfaces which can act as additional relaxation sources. In an important example, this is exploited for getting information about the pore structure for liquids in porous media [28].

3.2. Diffusion

While relaxation is sensitive to mobility on a molecular scale and is influenced by a combination of effects of the immediate molecular environment, the diffusion coefficient directly describes the translational motion of fluids. In liquids it is approximately inversely proportional to the bulk viscosity, but is reduced if geometrical restrictions exist. In a pulsed-field gradient NMR experiment, the self-diffusion coefficient D is not determined directly, but the mean-square displacement of the spins, which is related, in turn, to the diffusion coefficient by Einstein's equation:

$$<\boldsymbol{r}^2> = 2d\,D\Delta \tag{16}$$

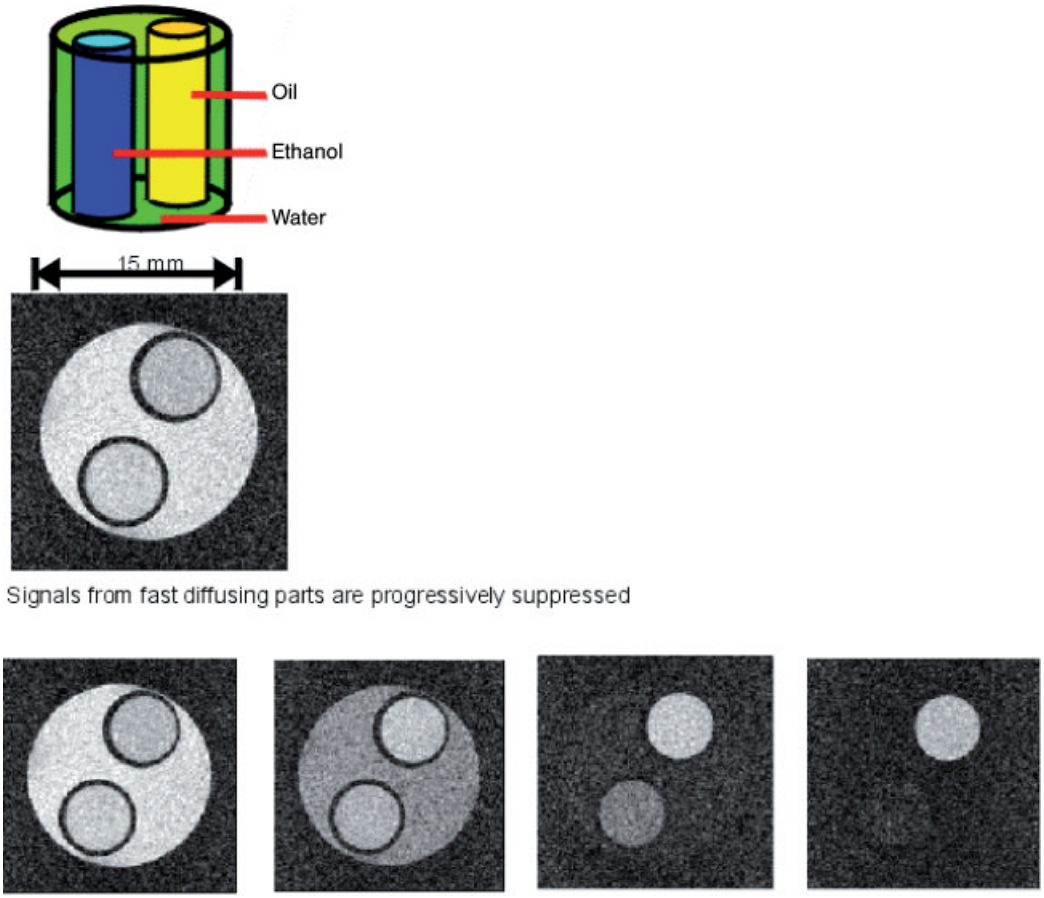

Figure 13. Diffusion-weighted images of three liquids with different self-diffusion coefficients. As the strength of the bipolar gradient is increased from left to right, the signal from the liquid with the largest diffusion coefficient (in this case, water) decays fastest. The initial image represents the ^{1}H spin density of the liquids. (Reproduced with permission from U. Scheler, Leibniz Institute for Polymer Research Dresden.)

Here, Δ is the time used for the encoding which is achieved by a bipolar gradient applied during a pulse sequence and acting on the transverse components of magnetization $M_{x,y}$, and d is the dimension in which displacements are measured ($d = 1$ for diffusion measurements along one axis). The signal intensity is then damped by a factor

$$S(\boldsymbol{q}) = \exp\left[-(2\pi)^2 \boldsymbol{q}^2 D\Delta\right] \tag{17}$$

This factor can be multiplied into an equation such as Equation 15 where D can also be assumed to be dependent on position, $D(x,y,z)$. The vector $\boldsymbol{q}$ is the area of one of the lobes of the bipolar gradient and is thus mathematically equivalent to the definition of $\boldsymbol{k}$ (see Section 5.1 for a more detailed description).

Image contrast arises because volume elements containing fluid with a larger diffusion coefficient will have a lower intensity. Fluid distributions with different viscosities, or fluids in materials with geometrical confinement, can be visualized by diffusion contrast methods. (see Fig. 13). In the latter case, restricted diffusion is observed if the mean-square displacement during Δ is comparable to the restriction size. Typical values of encoding times Δ in these types of experiments are on the order of 0.1 – 100 ms, and displacements are in the range of 0.1 – 10 µm for liquids.

The vectorial dependence of the displacement can be exploited to determine anisotropies of diffusion processes by repeating the experiment with bipolar encoding gradients in differ-

ent directions. A prominent example from medical diagnosis is diffusion tensor imaging in the brain which is called "fiber tracking": the diffusion coefficient component along the nerve bundles is larger than the component perpendicular to it due to the geometrical restriction of the latter.

3.3. Chemical Shift Imaging

Unique to NMR, and therefore the most powerful "contrast" of MRI, is its capability to differentiate the chemical constitution of the material via the chemical shift dispersion of the nuclei to be examined. In conventional NMR spectroscopy, the FID is acquired in a magnetic field of maximum homogeneity by subsequent Fourier transformation, which allows one to distinguish chemical shift and hyperfine coupling effects down to frequency differences of less than 1 Hz, requiring field homogeneities of 10^{-10} or better [1]. This immediately prohibits the combination of frequency encoding with direct NMR spectroscopy; instead, phase encoding methods must be applied which are characterized by signal acquisition in a homogeneous field.

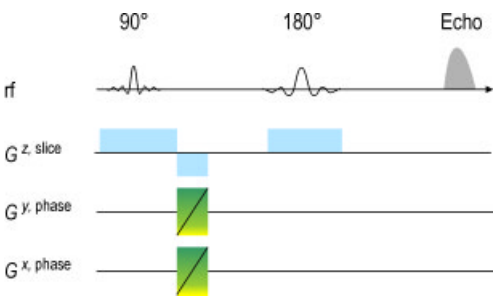

Figure 14. Chemical shift imaging (CSI) sequence with slice selection along z, and phase encoding along x and y spatial dimensions. The NMR signal is acquired in the absence of gradients. The sequence prior to signal acquisition can be regarded as a filter delivering chemical shift spectra for each of the image pixels within the slice.

The most prominent pulse sequence to obtain an image in which each pixel contains a full chemical shift spectrum is called chemical shift imaging (CSI). Figure 14 illustrates a CSI imaging sequence in which a slice dimension and two orthogonal dimensions of spatial encoding are defined before acquiring the signal in the absence of a magnetic field gradient. Each NMR spectrum obtained corresponds to a single point in 2D k space and delivers full chemical shift information for each pixel in the image after Fourier transformation, but it still involves averaging over the thickness of the excited slice. However, one can even replace the slice gradient by one further variable phase-encoding gradient along the remaining axis. One then performs a truly four-dimensional experiment, since it relates the signal in the (k_x,k_y,k_z,t) space to the information in the (x,y,z,ω) space by a four-dimensional Fourier transformation. CSI is therefore the most informative, but also the most time-consuming approach.

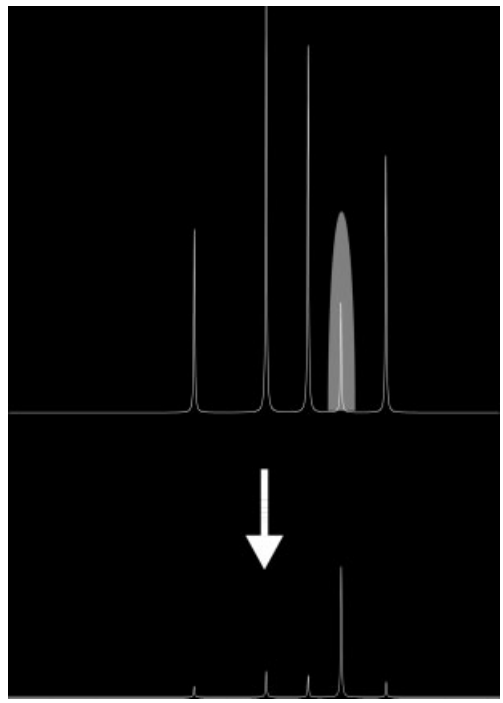

Figure 15. The effect of an rf pulse with finite bandwidth (indicated by the shaded area) on the NMR spectrum. Relative to the undisturbed spectrum (top), the spectrum following the frequency-selective pulse consists of a line within the bandwidth, and possibly of remainders of the other lines at much reduced intensity.

Instead of acquiring the entire chemical shift information as an extra dimension, one can reduce one dimension by utilizing frequency-selective rf pulses which only excite a nar-

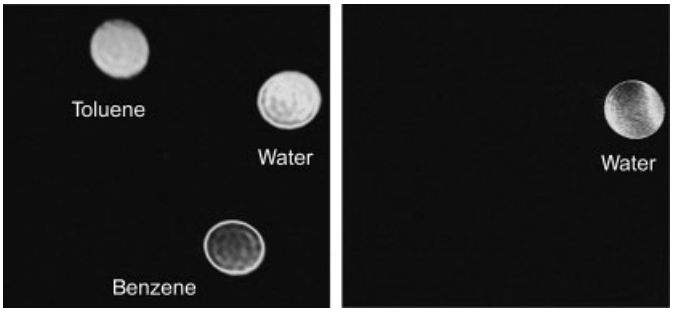

Figure 16. Imaging of ^{129}Xe using a) regular imaging and b) chemical shift selective imaging sequence along the cross section of three NMR tubes containing toluene, water, and benzene. All tubes contained dissolved hyperpolarized ^{129}Xe in the solvent. Selective excitation selected only the desired substance as is demonstrated for ^{129}Xe dissolved in water.

row bandwidth of interest. Here, the frequency-selective pulse acts as a filter prior to the imaging sequence and imposes a contrast on the image reflecting the chemical composition of the sample. Figure 15 symbolically describes the effect of such a pulse under realistic conditions: out of a spectrum of several lines, the selective pulse excites only spins with frequencies within its bandwidth, but often the remaining lines are not perfectly suppressed, a problem which requires more elaborate sequences in those frequent cases where a diluted component has to be highlighted with respect to a strong solvent background.

Figure 16 illustrates the effect of chemical shift selective imaging on three NMR sample tubes containing toluene, water, and benzene placed inside an NMR imaging probe. Hyperpolarized ^{129}Xe (see Section 6.3) was inserted into all three tubes and subsequently the ^{129}Xe NMR signal was obtained. ^{129}Xe has a large chemical shift range so that different solvents such as toluene versus benzene lead to a 10 ppm shift which renders chemical shift selective imaging very easy. In a), a regular spin warp imaging sequence was applied and in b), only the ^{129}Xe dissolved in water was selectively irradiated in order to obtain an image only from the water-containing part of the sample. Note that in principle, the same procedure can be carried out by utilizing the ^{1}H signals of the solvents themselves, but the much smaller chemical shift range of the hydrogen nuclei makes this a less straightforward experiment.

Another approach to obtain spatially selective chemical shift information is, instead of obtaining the entire image, to select only the voxel of interest of the sample and to record a spectrum. This method called volume-selective spectroscopy (VOSY) [29] is a 1D NMR method and is accordingly fast compared to a 3D sequence such as CSI. In Figure 17, a VOSY sequence based on a stimulated echo sequence is displayed, where three slice-selective pulses excite coherences only inside the voxel of inter-

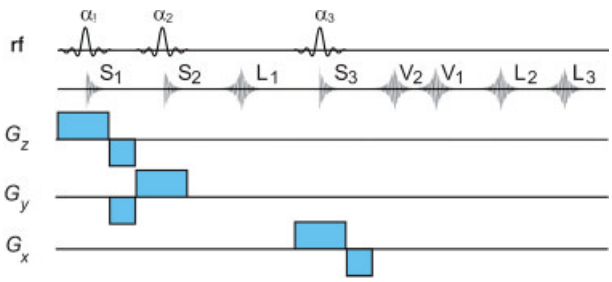

Figure 17. Volume-selective spectroscopy (VOSY) sequence based on a stimulated echo sequence using three selective pulses of arbitrary flip angle (α_1, α_2, α_3). These soft pulses are used for volume selection along x, y, and z. The echoes V_1 and V_2 are generated as an effect of all three rf pulses and can be exploited directly for localized spectroscopy of the selected volume, while the echoes L_1, L_2, and L_3 are generated by different combinations of two rf pulses and represent the averaged spectra within three orthogonal lines crossing at the center voxel.

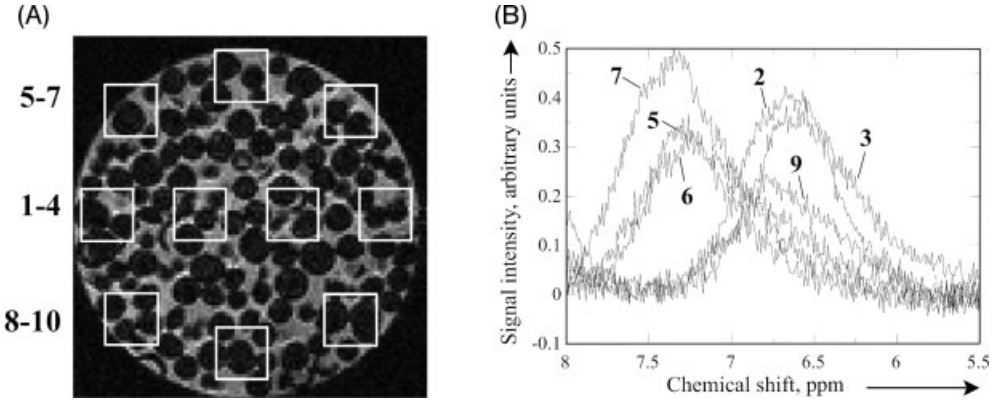

Figure 18. Volume-selective spectroscopy (VOSY) within a fixed bed of 20 mm diameter. A) Two-dimensional spin density image within a 0.5 mm thick slice near the center of the fixed bed. Selected volumes of 1.5 × 1.5 mm are indicated. B) Section of ^{1}H spectra covering the chemical shift range of the interparticle OH line; spectra were taken in the selected volumes shown in a); the acquisition time for each spectrum was 3 min. (Reproduced from [30] with permission).

est. The offset frequencies of the slice-selective pulses define the location of the voxel. Along the receiver axis all FIDs and echoes created by this stimulated echo sequence are displayed; they contain signals from different regions of the sample. Figure 18 demonstrates how such a sequence can be employed for monitoring an esterification reaction: the ^{1}H chemical shift of an OH line is a function of the pixel coordinates for different positions within a resin bed of 20 mm diameter. This chemical shift is a function of the relative amounts of acetic acid, methanol, and water because hydroxyl protons undergo fast exchange between these three molecules, which results in a weighted average of the corresponding spectral frequencies. The local concentration of acetic acid is thus directly obtained from the chemical shift, and the extent of conversion can be followed with spatial resolution by acquiring a series of VOSY spectra.

3.4. Spin Density and Parameter Images

Discussion about contrast in NMR imaging can be continued extensively. Merely the most basic contrast parameters for NMR imaging have been mentioned. There are endless varieties of "filter" sequences that can be set prior to an imaging sequence to specifically weight or select the NMR signal according to the nature of the physicochemical "filter". For instance, contrast representing dipolar coupling of various strength, order, and alignment, viscosity, and temperature can be created, to name just a few examples.

While filters can be applied at will, and serve for generating images weighted with a particular property, the two contrast parameters T_1 and T_2 cannot be "switched off", since the execution time and repetition time in any imaging experiment are finite. Only with knowledge of the relaxation properties is it possible to reconstruct the "true" spin density distribution. From the structure of Equation 15 it is clear that this can only be achieved by a comparison of signal intensities with different values of TE and TR, which requires at least three combinations of these waiting times. The signal intensity can then be extrapolated to TE = 0 and TR = ∞. This is based on the assumption that relaxation can be described by an exponential relation; otherwise, reconstruction requires some assumptions and, to minimize errors, short TE and long TR within the experimentally accessible range are preferred.

From comparing two or more images with a different weighting, it is just one more step to generating a proper parameter image. The difference is as follows: a weighted image still contains the spin density in the signal amplitude and represents a classical contrast image, whereas a parameter image only contains the values of the pure parameter (e.g., diffusion coefficient, T_1, T_2) in each pixel. For example, one can obtain two images with different TE while keeping all

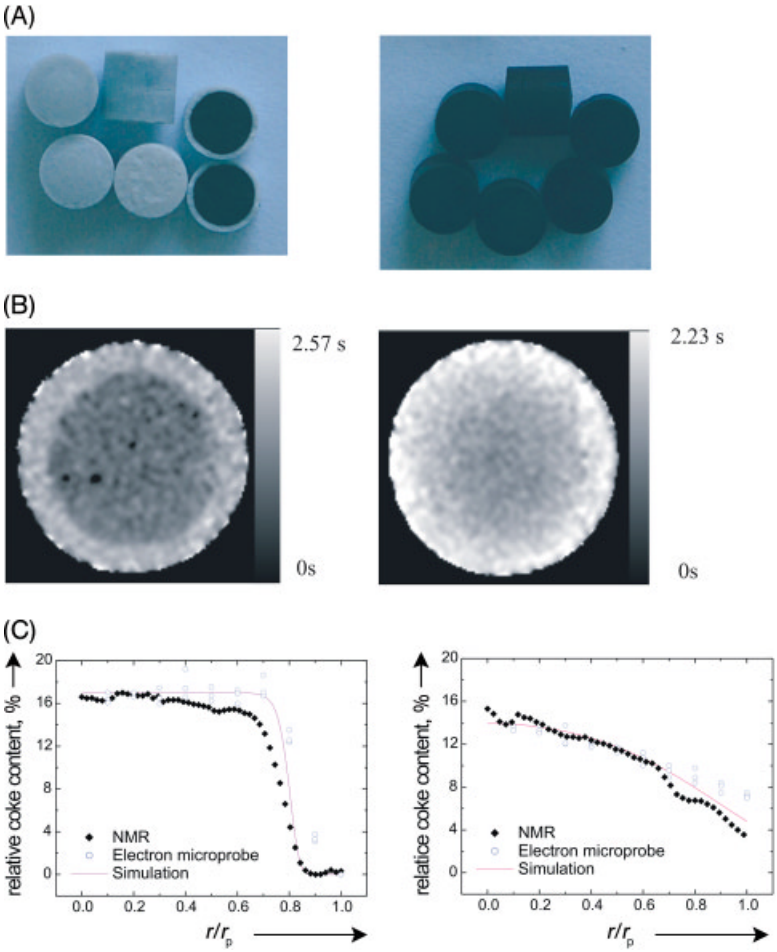

Figure 19. Partially regenerated, coked Al_2O_3 catalyst samples with the same residual coke content of 7.65 %. Left: regenerated at 550°C; right: regenerated at 400°C. A) Optical photographs of cut samples; B) NMR images with the relaxation time T1 as contrast parameter; C) Radial coke profiles determined from B), compared to electron microprobe measurements and simulated distributions.

other variables unchanged. Dividing the images pointwise

$$\varrho\,(x,y,z)\ \exp\left[-\mathrm{TE1}/T_2\,(x,y,z)\right] \\ /\varrho\,(x,y,z)\ \exp\left[-\mathrm{TE2}/T_2\,(x,y,z)\right] \\ = \exp\left[(\mathrm{TE2}-\mathrm{TE1})\ T_2\,(x,y,z)\right] \tag{18}$$

renders an image that contains the relaxation times averaged within each volume element. Figure 19 shows an example from chemical engineering: alumina pellets that have been used in hydrocarbon conversion accumulate coke residues on their internal and external surfaces which partially block the pore structure. Regeneration, i.e., removal of the coke, depends on the process temperature and is more homogeneous at lower temperatures. The coke-free zones at an intermediate stage of regeneration are easily visible, but the correlation of the relaxation time T_1 with coke content allows the reconstruction of a coke concentration profile across the pellet.

4. Fast Imaging Techniques

To speed up imaging, waiting times for retrieving full equilibration of magnetization before each subsequent encoding step are prohibitive. It therefore appears that relaxation weighting becomes unavoidable if fast imaging experiments are to be performed. However, there are fundamentally different approaches that, with modern

hardware, eventually reduce the acquisition time of three-dimensional images down to about one second. One can distinguish four strategies to achieve this goal: successive excitation of different partial volumes of the sample, successive excitation of fractions of the total magnetization by short rf pulses, generation and repeated encoding of a train of spin echoes, and the same approach for gradient echoes.

4.1. Multislicing Methods

One rather straightforward way to minimize image acquisition time is to expand the method of slice selection introduced in Section 2.3. Not only the thickness but also the position of the excited slice can be manipulated by varying the carrier frequency of the soft rf pulse. The same kind of 2D imaging experiment can thus be performed as was shown in Figure 7, but with each encoding step performed within a slice at a different position in the object. A series of slices can be defined in such a way that the total volume is covered, and 2D images within each slice can be combined to a 3D image of the object. Therefore, there is no need for a waiting time after the end of data acquisition, and the next encoding step can be applied immediately afterwards but for a different slice. The first slice can again be addressed after a total time of approximately 5 T_1 during which each slice has been excited once. The principle of multislicing is shown in Figure 20.

4.2. Nonequilibrium Methods

Another means for overcoming the dilemma of long experimental times is to employ small flip angles for the pure phase-encoding sequence. In the picture presented in Figure 1, transverse magnetization can be generated by an rf pulse of arbitrary length; out of the initial value M_0, a fraction $\sin\theta$ (where θ is the flip angle) is rotated into the transverse plane which is available for image encoding, and the remaining amount $(1-\sin\theta)$ is preserved. Because there is no longer the need to wait a multiple of the relaxation time T_1 for recovery of magnetization, much shorter repetition times can be used between the phase-encoding cycles. This method is termed fast low-angle shot imaging (FLASH) [31]. The fast repetitive excitation of the residual z magnetization causes signal depletion, but this is compensated to a certain degree by partial magnetization recovery due to spin – lattice relaxation. A typical flip angle (θ in Fig. 21) would be 5°. In reducing the flip angle, the total image acquisition time of FLASH is shortened at the expense of the image signal magnitude. Thus, it is possible to trade off imaging speed and signal-to-noise ratio in a controllable manner. Once an equilibrium between excitation and relaxation has been reached, the optimum flip angle is expressed by the so-called Ernst angle relation:

$$\cos\theta = \exp\left(-\mathrm{TR}/T_1\right) \tag{19}$$

Figure 20. Schematic of multislicing. By shifting the carrier frequency of the soft pulse, slices at different positions of the object are excited. The first soft pulse selects slice 1, and a line in 2D k space within this slice is acquired by setting a particular value of the phase gradient. This encoding is then repeated for slice 2, and so forth. At the last slice, the magnetization in slice 1 has fully relaxed and the next phase-encoding step can be carried out at this position. After the end of the experiment, data must be rearranged so that individual reconstruction of the 2D images in each slice is possible. Note that small gaps between slices may be used to avoid cross-talk; alternatively, one can avoid scanning of successive slices by choosing, for instance, the order 1 – 6 – 2 – 7 – 3 – 8 – 4 – 9 – 5 – 10.

A fast imaging technique particularly tailored for examining solid samples with extremely short T_2^* relaxation time is to employ pure phase encoding, acquire a single point instead of a full FID, and leave the phase encoding gradients

turned on during the entire sequence in order to save the time normally invested into switching and stabilizing of the gradient. Under such conditions, neither the generation of an echo nor the execution of a slice-selective pulse is feasible because of the fast coherence loss of the magnetization. The basic scheme (single-point imaging, or SPI) is sketched in Figure 22A, where gradients along all three coordinates are varied independently of each other.

The modification of the weight function to access the $\boldsymbol{k}$ vector can also be realized via a step-like or continuous change of the gradient function. The acquisition of only one point per phase encoding cycle, i.e., rf pulse and gradient step (single-point ramped imaging with T_1 enhancement, or SPRITE, [32]) with minimized dead time and waiting time between rf pulse and acquisition makes this sequence fast and robust. A two-dimensional realization of this method is shown in Figure 22B, where both gradients are modulated in a sinusoidal fashion in order to avoid large sudden gradient changes and to reduce total energy dissipation; by this method, k space is scanned in a spiral originating at the origin. An experimental result of 2D and 3D SPRITE experiments with optimized k space coverage is shown in Figure 23.

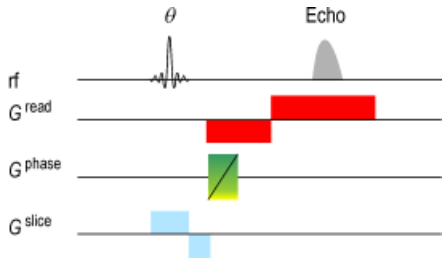

Figure 21. Gradient-echo pulse sequences based on low flip angles, known by the acronym FLASH (fast low-angle shot) imaging [31]. The initial rf pulse defined the excited slice; the coherence of the signal is then re-established by a gradient echo in "read" direction. For each repetition, a different value of the phase-encoding gradient is set, so that a fast 2D image within the defined slice can be acquired.

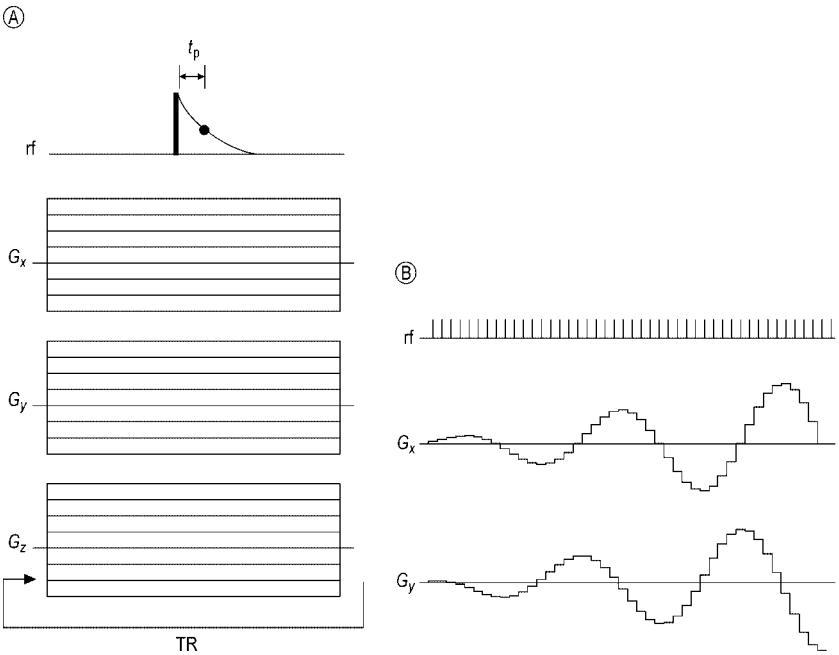

Figure 22. A) Basic principle of the single-point imaging (SPI) scheme. Gradients in three orthogonal directions are varied independently of each other. One single point is acquired for each combination of gradient values, with the encoding time tp between rf pulse and acquisition event ideally kept as short as possible, taking the dead time of the receiver into account. Gradients are switched on a finite period before the rf pulse to avoid detrimental effects of gradient rise time and eddy currents. B) 2D Spiral-SPRITE method. Gradients are kept constant only during rf pulse application and data acquisition. Because of the sinusoidal modulation of the gradient intensities, the change of the gradient values between successive acquisitions is minimized, allowing for shorter prepulse delays. (Reproduced from [33] with permission).

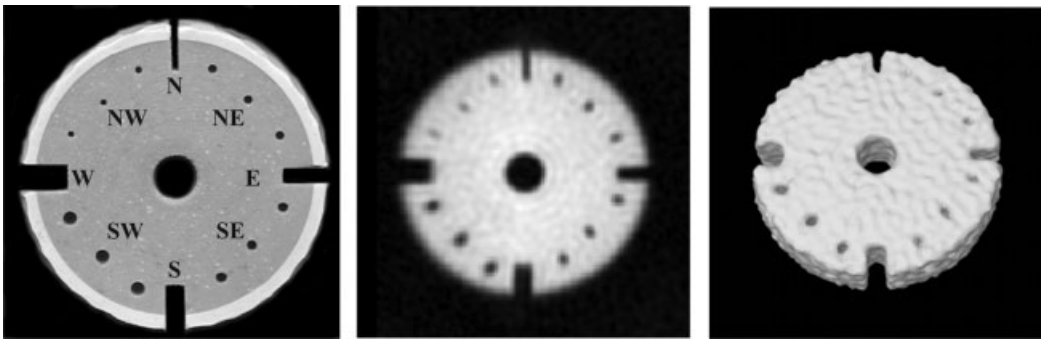

Figure 23. SPRITE experiments on a cross-linked *cis*-polybutadiene disk manufactured for demonstrating spatial resolution. Left: sample geometry; outer diameter: 43 mm, outer ring (invisible in experiment) thickness: 2.5 mm; diameter of the holes: 1.0, 1.25, 1.75, and 2 mm (NW to SW). The material has a short T_2 relaxation time of 560 ms that excludes conventional imaging based on spin echoes. Center: 2D SPRITE image with 64 × 64 pixels, experiment time 16 s. Right: 3D SPRITE image with 64 × 64 × 64 voxels, experiment time 112 s. The images were acquired with an optimized *k* space trajectory discussed in [34] (reproduced with permission from the authors).

4.3. Multiple Spin-Echo Sequences

The fundamental idea behind the basic spin-echo sequence shown in Figure 1 (and combined with position-encoding magnetic field gradients in Figs. 3, 7, and 8) can be extended to produce a series of echoes. Under the influence of local field inhomogeneities or chemical shift differences, the spin isochromats maintain their individual differences in Larmor frequency and are beginning to dephase again from the moment of the echo maximum. As was pointed out above, the second half of the echo is equivalent to an FID, so it can be taken as a new starting point of a pulse sequence. Another 180° pulse has the effect of refocusing the dephasing transverse magnetization, and generates a second echo. A series of 180° pulses can thus be employed to produce a train of hundreds or thousands of echoes, each of them reduced in amplitude by spin – spin relaxation with a time constant T_2 to a fraction $\exp(-t/T_2)$ of the FID amplitude, where t is the total elapsed time since the first 90° rf pulse. With appropriate phase cycling of the pulses, this method is called the Carr–Purcell–Meiboom–Gill (CPMG) sequence and constitutes the standard sequence for measuring T_2.

Rather than just determining the amplitude of each echo, one strategy for performing fast imaging is to encode position information into each echo in a different way, e.g., by applying phase gradients of different strength between the rf pulses and reading the signal intensity with a frequency encoding gradient. Depending on the duration of signal acquisition and the value of T_2, all necessary phase encoding steps, or at least

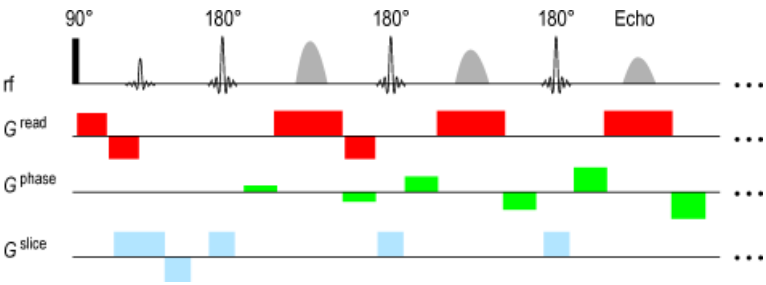

Figure 24. RARE sequence for fast imaging. The formation of the first spin echo is similar to a conventional imaging sequence such as that shown in Figure 8, but in this case all rf pulses are considered slice-selective. All further spin echoes are encoded by phase gradients of different strength. The CPMG form of spin echo is used to avoid the accumulation of flip-angle errors over the echo train. However, before the following echo can be acquired, the phase-encoding has to be "rewound" to undo the dephasing of the spins. Therefore, a phase encoding step of equal strength but opposite sign is applied after the completion of each data acquisition. If the entire *k* space data set is to be acquired with multiple RARE echo trains, the whole sequence is repeated with different phase-encoding steps until the entire *k* space is covered.

several of them, can be performed within a single echo train. One prominent sequence based on multiple spin-echoes that works according to this principle is called rapid acquisition with relaxation enhancement (RARE) [35]. The so-called turbo spin-echo (TSE) sequence is a commercial version of RARE. RARE is based on a CPMG echo train and is displayed in Figure 24. Acquisition of the first echo signal is conventional; it includes a phase-encoding and a read gradient. However, before the second echo can be acquired, the phase encoding must be "rewound" to undo the dephasing of the spins. Rewinding is realized by a phase-encoding step of equal strength but opposite sign applied after completion of data acquisition. At this instance, the total phase shift of the isochromats due to position encoding is zero, so that each 180° pulse acts on the system with equal properties. The whole sequence is repeated after a recovery time with different phase encoding steps. Within one echo train, a portion of k space is covered. So, if l echoes are acquired per scan, and n scans are carried out with varying phase-encoding gradients, a total of $l \times n$ number of lines in k space will be collected. In other words, each echo is responsible for a different portion of k space. One drawback is that different lines of k space will have different T_2 weighting because they are acquired at different echo times. Depending on the exact design of the sequence, one can assign any of the echoes to coincide with the center or outer part of k space, which then affects the contrast of the image.

The use of rewinding gradient pulses has made RARE rather robust since, particularly in structured media, magnetic fields inside the sample are intrinsically inhomogeneous so that local shifts in resonance frequencies and thus phase errors are unavoidable. If this problem is less pronounced, a "blipped" version of a multiple spin-echo sequence can be employed, where the increase of the value of k in phase direction is achieved in a cumulative manner by applying the same, small PFG (so-called "blip") before each echo acquisition. Avoiding the rewinding gradient saves time in the execution of the pulse sequence. In a similar fashion, a number of other variations based on the principle of spin echoes have been suggested.

4.4. Multiple Gradient Echo Sequences

In the quest to reduce the experiment time for acquiring a series of echoes with different positional encoding, the necessary waiting times for generating spin echoes are a disadvantage. Likewise, a compromise must be made between image resolution and acquisition time – a high resolution requires the readout of many points during the acquisition period, so that acquisition time can eventually only be saved at the cost of limited resolution. The theoretically fastest NMR imaging sequence is defined in such a way that a maximum of the experimental time is used for gathering information, i.e. the final goal would be to achieve quasicontinuous data sampling. The use of gradient echoes instead of spin echoes almost matches this theoretical limit.

Figure 25 shows the principle of an imaging sequence based on multiple gradient echoes, called echo planar imaging (EPI) [36] because it was initially designed for the rapid acquisition of a 2D image of a slice that is defined by the

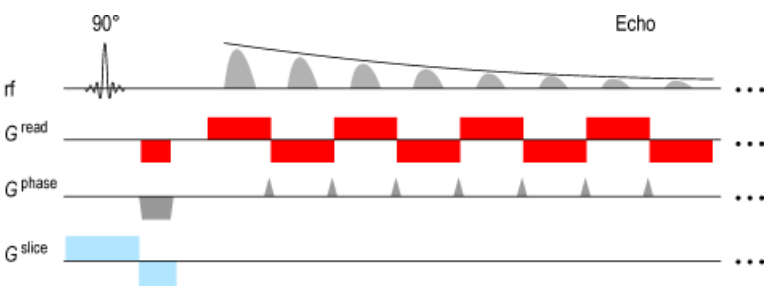

Figure 25. Echo planar imaging (EPI) pulse sequence. Gradient-echo-based multiple echoes are used for fast single-shot 2D imaging. Slice selection along G_s and frequency encoding along G_r is utilized. Phase encoding is realized by using short "blipped" gradient pulses along G_p.

initial soft 90° pulse. No further rf pulses are used throughout this sequence.

According to the above definition, EPI is the fastest pulse sequence available, offering acquisition times of 100 ms or less for 2D images. Its main disadvantage is the fact that the echoes are generated by the rephasing effect of the applied gradients themselves, which alternate in sign; however, the effects of background gradients and chemical shift differences are not refocused. EPI is best used with objects of rather homogeneous magnetic field distributions, and with spectra in which the chemical shift difference is negligible. The train of echoes and the image intensity itself are T_2^*-weighted. An EPI image is thus feasible only when T_2^* of the FID exceeds the total acquisition time of the echo train; this can serve as a first test of the applicability of EPI. In the version depicted in Figure 25, the variation of the phase direction of k space is achieved by small "blip" gradient pulses (see discussion in Section 4.3). The blips can be replaced by a constant, weak gradient that leads to a continuous accumulation of the signal phase, which is interrogated at the position of each echo; the definitions of "phase encoding" and "frequency encoding" then lose their meaning in the strict sense, since the complete signal evolution takes place in the presence of magnetic field gradients.

Again, there are a number of further modifications to EPI sequences. Common to all is the need for high-quality gradient coils and amplifiers that allow the quick change of large currents in a defined and reproducible way.

5. Measurements of Transport Processes

In many modern applications of medical and nonmedical MRI, the investigation of motion has become the focus of attention. With comparatively little modification to the basic pulse sequences, motion can be detected by standard MRI hardware, while many other, classical techniques are only capable of generating still images. There are again two fundamentally different approaches to NMR imaging of motion: first, a quick succession of images acquired with the methods discussed in Chapter 4 is suitable to visualize transient phenomena on timescales of tens of milliseconds; these so-called cine methods have successfully been used for generating movies of, for instance, beating hearts in animals. This application can be regarded as a straightforward extension of pure imaging.

It is, however, also possible to encode parameters such as velocity or diffusion coefficients into an image, without the need to acquire two different images sequentially. The encoding takes place by pulsed field gradients of suitable properties. In the following three sections, the fundamental relationships of these dedicated methods are summarized. A different approach, exploiting various aspects of time-of-flight effects, is briefly discussed in Section 5.4.

5.1. q Space

The $\boldsymbol{q}$ vector was already introduced in Equation 17 in the context of diffusion measurements. More generally, it is used in experiments in which motion is to be quantified. In Figure 26, the simplest pulse sequence for determining motion by NMR is shown.

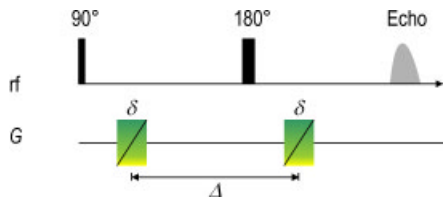

Figure 26. Basic pulse sequence for detecting motion. At the time of the spin echo, phase differences due to time-invariant field inhomogeneities and chemical shift differences are refocussed (see Fig. 1). The application of two pulsed gradients with the same area but opposite (effective) sign has no effect on static matter, but imposed a phase shift onto spins which have moved during Δ.

The transverse component of the magnetization $M_{x,y}$ accumulates all phase shifts that are imposed by position-dependent magnetic fields between the initial 90° rf pulse and data acquisition at the time of the echo. In this sequence, the magnetization experiences two pulsed gradients with equal area but opposite sign (note that all 180° pulses have the effect of alternating the sign of any PFG that was applied before the pulse). It can be regarded as a double encoding of posi-

tion, occurring at different times separated by a delay Δ.

The total phase can be written as

$$\begin{aligned}\phi(\boldsymbol{r}_1,\boldsymbol{r}_2) &= \gamma\delta(\boldsymbol{g}_1\boldsymbol{r}_1+\boldsymbol{g}_2\boldsymbol{r}_2)\\ &= 2\pi(\boldsymbol{k}_1\boldsymbol{r}_1+\boldsymbol{k}_2\boldsymbol{r}_2)\\ &= 2\pi\boldsymbol{k}_1(\boldsymbol{r}_1-\boldsymbol{r}_2)\\ &= -2\pi\boldsymbol{q}\,\boldsymbol{R}\end{aligned} \tag{20}$$

where $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ are the positions at the time of the first and the second gradient pulse, respectively. Because of the interdependence of the gradients, only one variable wave vector remains, and it determines the phase shift proportional to the displacement $\boldsymbol{R}$. Note that, by applying a matched pair of gradients, the information about position is totally removed and is replaced by a measure of displacement of spins at any (undefined) position within the sample.

In order to symbolize that displacements are encoded rather than positions, the corresponding wave vector is renamed $\boldsymbol{q}$. While it is obvious that $\boldsymbol{q} = \boldsymbol{k}_1 = -\boldsymbol{k}_2$, a change of variable names often facilitates the discussion of more complex encoding schemes when distinguishing whether position or displacement ought to be encoded. In most, but not all, recent works, $\boldsymbol{q}$ is used for the Fourier conjugate variable to $\boldsymbol{R}$.

The signal, expressed as a function of $\boldsymbol{q}$, can now be written as

$$S(\boldsymbol{q}) = \int \bar{P}(\boldsymbol{R},\Delta)\,\exp[i2\pi\boldsymbol{q}\,\boldsymbol{R}]\,\mathrm{d}\boldsymbol{R} \tag{21}$$

Equation 21 is formally equivalent to Equation 10 defining position encoding, but it contains as the kernel the distribution function of displacements $\int \bar{P}(\boldsymbol{R},\Delta)$. This function is also called the *propagator* and it should, if correctly used, include one further variable, the encoding time Δ between the gradient pulses over which the displacement $\boldsymbol{R}$ is determined. It is important to take Δ into consideration, because in many real situations not only the average displacement but also the shape of the propagator will depend on the length of the interval during which it is being measured.

The bar above the symbol of the propagator denotes an ensemble average over the whole sample, or a selected volume of it; the position information is lost. The averaging is contained in the formal description of the propagator, which can be written as

$$\bar{P}(\boldsymbol{R},\Delta) = \int \varrho(\boldsymbol{r}_1)\,P(\boldsymbol{r}_1|\boldsymbol{r}_1+\boldsymbol{R},\Delta)\,\mathrm{d}\boldsymbol{r}_1 \tag{22}$$

In this expression, the propagator is decomposed into the *spin density* at the initial position $\varrho(\boldsymbol{r}_1)$ and the *conditional probability* that a particle moves from $\boldsymbol{r}_1$ to a position $\boldsymbol{r}_1+\boldsymbol{R}$ during the interval $\Delta P(\boldsymbol{r}_1|\boldsymbol{r}_1+\boldsymbol{R},\Delta)$. Note that neither of these functions can be directly retrieved because they are "hidden" in the integral. The propagator experiment is one-dimensional since it involves the simultaneous variation of both gradient pulses; however, by making use of more complex and higher dimensional experiments, conditional probability functions such as that mentioned here can indeed be determined [18, 37].

Just as the spin density distribution $\varrho(\boldsymbol{r})$, and hence the projection of the investigated object, is obtained in phase encoded imaging, $\bar{\boldsymbol{P}}(R,\Delta)$ can be reconstructed by applying the inverse Fourier transformation to $S(\boldsymbol{q})$, where one has to scan $\boldsymbol{q}$ space by varying the strength of the gradient pair. The resulting propagator might not be as intuitive as the function $\varrho(\boldsymbol{r})$, but it is a valuable tool to check distributions of displacements or velocities by comparing them with particular models. Some classical examples are given here.

Perhaps the simplest flow problem is that of laminar flow along z through a cylindrical pipe of radius r_0. For this so-called Poiseuille flow, the axial velocity v_z depends on the radial coordinate r as $v_z(r) = v_{\max}\lfloor 1-\left(\frac{r}{r_0}\right)^2\rfloor$, which is a parabolic distribution with the maximum flow velocity in the center of the pipe and zero velocity at the wall. The distribution function of velocities is obtained from equating $\int P(r)\,\mathrm{d}r = \int P(v_z)\,\mathrm{d}v_z$ and the result is that $P(v_z)$ is a constant between 0 and $v_{\max}$. The propagator has the same shape, as it measures the distribution of downstream displacements, $Z = v_z\Delta$. Often, the name propagator is used synonymously for the distribution of velocities; this is justified in this and similar cases, but not in general.

Figure 27 compares a sketch of the parabolic velocity profile with an example of an experimentally obtained propagator $P(Z)$ for flow in a pipe. It turns out that for this simple case, the shape of the distribution function is very sensitive to imperfections in the flow, which makes determination of the propagator a suitable method to assess the quality of laminar flow

conditions. For instance, at a closer look, the rectangular shape of the propagator is broadened by self-diffusion that takes place both parallel and perpendicular to the flow direction, and thus depends on the measurement time: for very short Δ, self-diffusion contributes a considerable broadening while for long Δ, radial Taylor dispersion leads to a change of flow velocities, and also smoothes the edges of the propagator. It has been found that subtle changes in propagator shape can be distinguished far better than direct determinations of velocity differences at a given point.

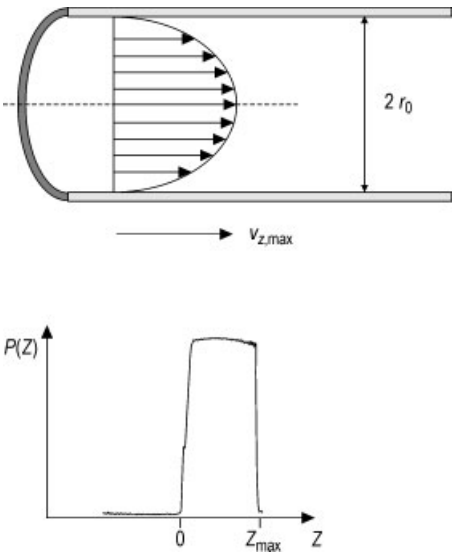

Figure 27. Left: schematic plot of the distribution of flow velocities v_z for laminar flow of a Newtonian fluid in a circular pipe; the maximum value of the velocity, which occurs in the center of the pipe, is drawn for comparison. Right: experimentally obtained propagator of water flowing under laminar conditions in a circular pipe; the probability density of displacements is constant between 0 and $Z_{\max} = v_{z,\max}\Delta$ where Δ is the encoding time of the experiment.

In Figure 28, the propagator of free self-diffusion is shown. Actually, self-diffusion was the first case for which the propagator terminology was discussed in the NMR context [38]. The propagator of free diffusion is of Gaussian shape; the mean-squared displacement of spins or molecules follows the same Gaussian evolution as the concentration of a pointlike source. It is of the form

$$\bar{P}\left(\boldsymbol{R},\Delta\right) = \left(4\pi D\Delta\right)^{-3/2} \exp\left[-\boldsymbol{R}^2/6D\Delta\right] \tag{23}$$

and is identical in all three directions. The result of an NMR experiment with an encoding wave vector $\boldsymbol{q}$ is as follows:

$$S\left(\boldsymbol{q}\right) = \exp\left[-(2\pi)^2\boldsymbol{q}^2\,\boldsymbol{R}^2/2\right] \tag{24}$$

This is again a Gaussian function. By using Einstein's equation (Eq. 16) one obtains the signal decay due to diffusion that was employed above (Eq. 17).

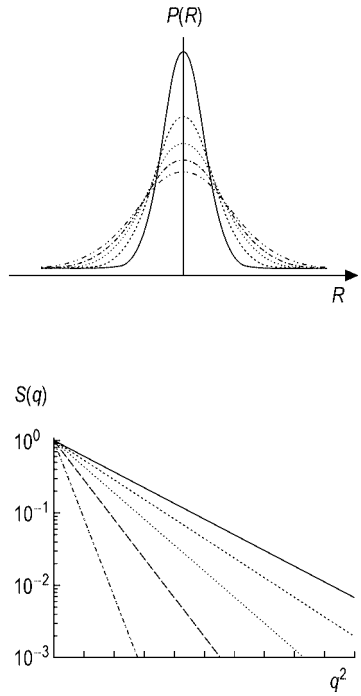

Figure 28. Top: propagator for unrestricted self-diffusion. The propagator $\bar{P}\left(\boldsymbol{R},\Delta\right)$ is shown for increasing encoding times Δ and becomes broader with increasing Δ, while its intensity at zero displacement is reduced due to requirement that the area remains normalized to unity. Bottom: signal function S($\boldsymbol{q}$) as would be obtained in an experiment, plotted semi-logarithmically against $\boldsymbol{q}^2$. In this representation, the slope of the decaying function is equal to $(4\pi)^2\Delta D$, so that the diffusion coefficient D can be obtained directly from comparing at least two measurements taken at different values of $\boldsymbol{q}$.

The Gaussian shape of a propagator leads to considerable advantage as far as NMR experiments are concerned: it is not necessary to resolve the complete shape of the propagator, which is fully described by three parameters: amplitude, center position, and width. While the amplitude is usually irrelevant and the average

displacement of free diffusion is zero (which has the consequence that the signal function $S(\boldsymbol{q})$ contains no complex phase factor), only the width of the propagator is of interest. One can obtain this quantity by performing a simplified experiment. If one plots $S(\boldsymbol{q})$ as ln S versus $\boldsymbol{q}^2$, a straight line is obtained (right part of Fig. 28). The slope of this line is proportional to the self-diffusion coefficient D and can — at least in principle — be obtained by measuring at just two different $\boldsymbol{q}$ values.

Restricted diffusion, correlated motion of spins, or any deviation from a "free" behavior of the molecules will result in a non-Gaussian propagator shape. A wide range of studies have dealt with such problems during the last two decades, and NMR has turned out to be the method of choice for quantifying restricted-diffusion phenomena such as liquids in porous materials and dynamics of entangled polymer molecules; for an overview of methods, see [39, 40]. For pure displacement measurements like this, the term "q space imaging" has been coined, which is somewhat misleading, since it describes the visualization of statistical distributions of displacements, instead of real images.

5.2. Motion Encoding and Moments

From the discussion of bipolar gradients in Section 3.2, it becomes clear that NMR pulse sequences can be designed which contain information about position, velocity, and potentially other parameters that are related to motion [41, 42]. The individual tools that PFGs provide for the encoding are best explained by extending the relationship for the accumulated phase shift in the presence of a magnetic field gradient (Eq. 9) towards a general expression that allows for variation of gradients and positions with time:

$$\phi(\boldsymbol{r},t) = \int_0^t \omega(t') \cdot \boldsymbol{r}(t')\,\mathrm{d}t' = \int_0^t \gamma \boldsymbol{g}(t') \cdot \boldsymbol{r}(t')\,\mathrm{d}t' \tag{25}$$

If the position of a spin is expanded into a Taylor series:

$$\boldsymbol{r}(t) = \boldsymbol{r}_0 + \boldsymbol{v}_0 t + \frac{1}{2}\boldsymbol{a}_0 t^2 + \ldots \tag{26}$$

one obtains [18]:

$$\begin{aligned}\phi(\boldsymbol{r},t) &= \gamma \int_0^t \boldsymbol{g}(t') \cdot \left[\boldsymbol{r}_0 + \boldsymbol{v}_0 t' + \tfrac{1}{2}\boldsymbol{a}_0 t'^2 + \ldots\right] dt' \\ &= \gamma \left[\boldsymbol{r}_0 \int_0^t \boldsymbol{g}(t')\,\mathrm{d}t' + \boldsymbol{v}_0 \int_0^t \boldsymbol{g}(t')\,t'\mathrm{d}t' \cdot \right. \\ &\quad \left. \tfrac{1}{2}\boldsymbol{a}_0 \int_0^t \boldsymbol{g}(t')\,t'^2\,\mathrm{d}t' + \ldots\right] \\ &= \gamma\left[\boldsymbol{r}_0\boldsymbol{m}_0 + \boldsymbol{v}_0\boldsymbol{m}_1 + \tfrac{1}{2}\boldsymbol{a}_0\boldsymbol{m}_2 + \ldots\right]\end{aligned} \tag{27}$$

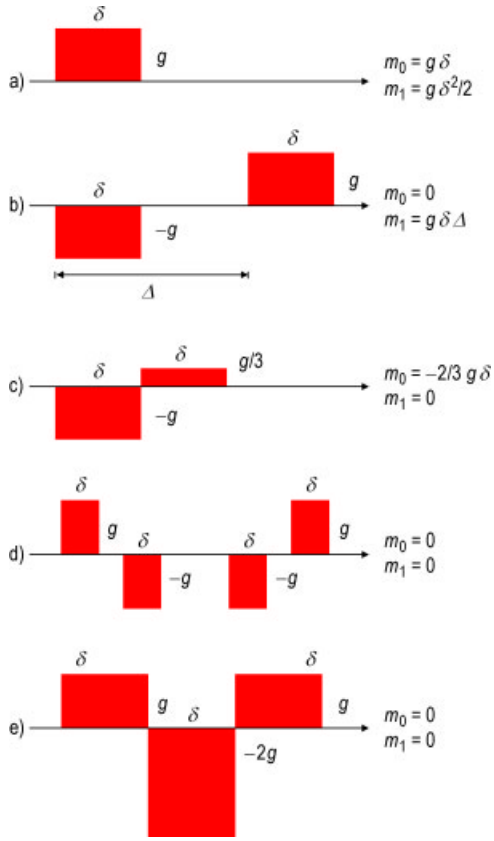

Figure 29. Typical pulsed gradient shapes commonly used for imaging or measurements of motion. a) A single gradient pulse, which is used for phase encoding of position, also has all other moments nonvanishing, so that the signal phase is affected by position, velocity, and acceleration. b) A bipolar gradient is the basic building block of most measurements of diffusion or flow; due to its vanishing zeroth moment, its application does not contain any position information. c) A phase-encoding gradient that is insensitive to motion needs a more complex shape; at least of two lobes are required. d) Two bipolar gradients arranged back-to-back generate an encoding that is affected neither by position nor by velocity, but higher order terms such as acceleration can be measured. The variant shown in e) is equivalent to d) but with the time intervals between pulses kept to a minimum. Note that in all cases, the effective gradient is shown; each 180° rf pulse applied anywhere within the sequence has the effect of reverting the sign of all gradient pulses that precede it so that the effective gradient becomes the negative equivalent to the physically applied gradient.

Terms of higher order than acceleration are usually not considered in NMR experiments. One can thus identify three integrals which are the respective Fourier conjugates to position $\boldsymbol{r}_0$, velocity $\boldsymbol{v}_0$, and acceleration $\boldsymbol{a}_0$:

The zeroth moment $\boldsymbol{m}_0$ was already mentioned briefly in Section 2.4. It is nothing else than the total area of the gradient (see Fig. 29a). If the area vanishes, the signal is independent of spin position. This is the case, for instance, for a bipolar gradient with positive and negative lobes of equal area (see Fig. 29b).

The first moment $\boldsymbol{m}_1$ encodes velocity. One can easily show that the bipolar gradient in Fig. 29b has a first moment of $\boldsymbol{m}_1 = \boldsymbol{g}\delta\Delta$, while $\boldsymbol{m}_0 = 0$. Symmetric gradient pairs are therefore the simplest way to measure velocities while introducing no effect of position on the signal, i.e., the velocity encoded by such a gradient pair alone represents all spins in the whole sample. If phase encoding is to be achieved without an additional influence by velocity, the conditions $\boldsymbol{m}_0 \neq 0, \boldsymbol{m}_1 = 0$ must be met simultaneously (see

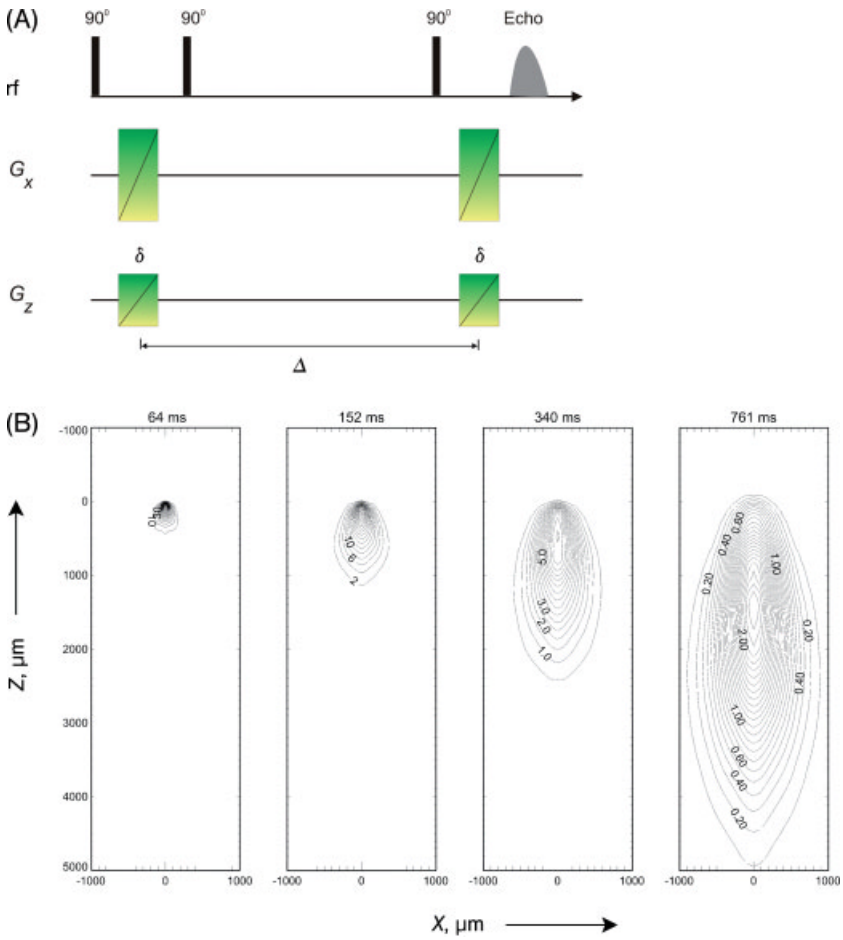

Figure 30. A) Pulse sequence for encoding of displacements in two dimensions. The sequence is based on the one-dimensional case (Fig. 26) but a stimulated echo is used instead of a spin echo; such a sequence adds further flexibility to the investigated timescale, since the encoding time Δ can be varied over a wider range compared to the spin-echo sequence, since the magnetization decays only with a time constant T_1 between the second and the third rf pulses. Both gradient pairs (in this case, G_x and G_z) are varied independently of each other. B) 2D propagators for water flowing through a random bed of monodisperse glass beads (600 mm diameter) at an average interstitial velocity of 2.5 mm/s for encoding times Δ indicated above the diagrams. The contour lines indicate the probability of displacements along the flow axis (Z) and perpendicular to it (X). (Reproduced from [46] with permission).

Fig. 29c, and the more detailed discussion of Fig. 33 in Section 5.3).

The second moment, $\boldsymbol{m}_2$, encodes acceleration. If only acceleration is to be measured, the conditions $\boldsymbol{m}_0 = 0, \boldsymbol{m}_1 = 0$ must be fulfilled. Figure 29d and e show two commonly used realizations. See [43 – 45] for applications in which acceleration is measured by NMR.

The discussion of "q space imaging" and similar concepts to measure displacements has been limited to one-dimensional experiments so far. However, just like in k space, experiments of higher dimension can be performed, although their interpretation requires more care.

The direct counterpart of 2D phase encoding of position is the 2D propagator, where gradient pairs are applied simultaneously but in orthogonal directions and varied independently of each other (see Fig. 30). What is obtained is a two-dimensional distribution function, $\bar{\bar{P}}\,(X,Z,\Delta)$, that can, for instance, be used to investigate how displacements along a flow axis Z are related to those perpendicular to the flow axis X. An example for flow of water through a random structure is shown in Figure 30B.

Of even more fundamental interest is probably the comparison of displacements in the same direction but at different times, an experiment called velocity exchange spectroscopy (VEXSY) [47]:

In a 2D VEXSY experiment, two independent pairs of bipolar gradient pulses (with separation Δ between the bipolar gradients) are used to encode displacements. These displacements (or velocities) are correlated at two different times separated by t_{m}, called the mixing time. This is in the form of a classical NMR spectroscopy exchange experiment, which is used to find correlations between features in the frequency spectrum by identification of cross-signals in a two-dimensional density representation $S(f_1, f_2)$ [1, 8]. Following the same principle that these off-diagonal cross-signals represent spins that are correlated and have changed or transferred their properties because of various coupling mechanisms during a mixing period or pulse, the off-diagonal cross-signal in a 2D VEXSY map represents spins that have changed their initial velocity after a mixing time of τ_{m}. The resulting function of a 2D VEXSY experiment is the joint probability density of finding displacements $\boldsymbol{R}_1$ during the first and $\boldsymbol{R}_2$ during the second interval, separated by the mixing time:

$$\bar{\bar{P}}_2\,(\boldsymbol{R}_1,\Delta;\boldsymbol{R}_2,\Delta;\tau_{\mathrm{m}}) = \bar{P}_1\,(\boldsymbol{R}_1,\Delta)\,\mathrm{P_V}\,(\boldsymbol{R}_1,\Delta|\boldsymbol{R}_2,\Delta;\tau_{\mathrm{m}}) \tag{28}$$

where the conditional probability $\mathrm{P_V}\,(\boldsymbol{R}_1,\Delta|\boldsymbol{R}_2,\Delta;\tau_{\mathrm{m}})$, describing the probability of displacement $\boldsymbol{R}_2$ given that it was in the first interval, can be interpreted more easily and is obtained by dividing the 2D function by the simple propagator $\bar{P}_1\,(\boldsymbol{R}_1,\Delta)$ [37].

A 2D VEXSY experiment can be based on a spin echo where transverse magnetization carries the information of the moving spins during τ_{m}, or on a stimulated echo where the longer lived longitudinal magnetization carries the information.

Figure 31 depicts one possible pulse sequence for a VEXSY experiment. Here, the first pulsed gradient pair (wave vector $\boldsymbol{q}_1 = \boldsymbol{k}_1 = -\boldsymbol{k}_2$) is placed between the initial 90° rf pulse and the first spin echo (not shown). The second

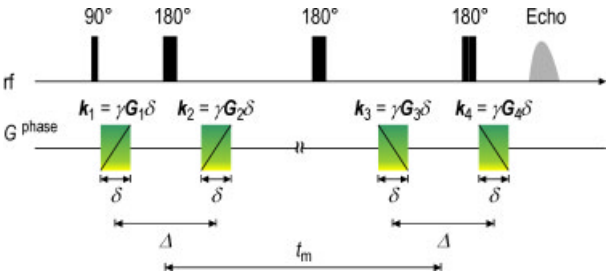

Figure 31. Spin-echo-based pulse sequence to encode velocity change. The gradients are stepped pairwise independently (2D VEXSY) or simultaneously (1D VEXSY). For a VEXSY experiment, k_1 to k_4 usually are applied along the same spatial direction. δ is the duration of each gradient pulse, Δ the separation between each pair of bipolar gradient pulses and tm the mixing time between the bipolar gradient pairs. The opposite polarity of the bipolar gradient pair is realized by an 180° inversion pulse.

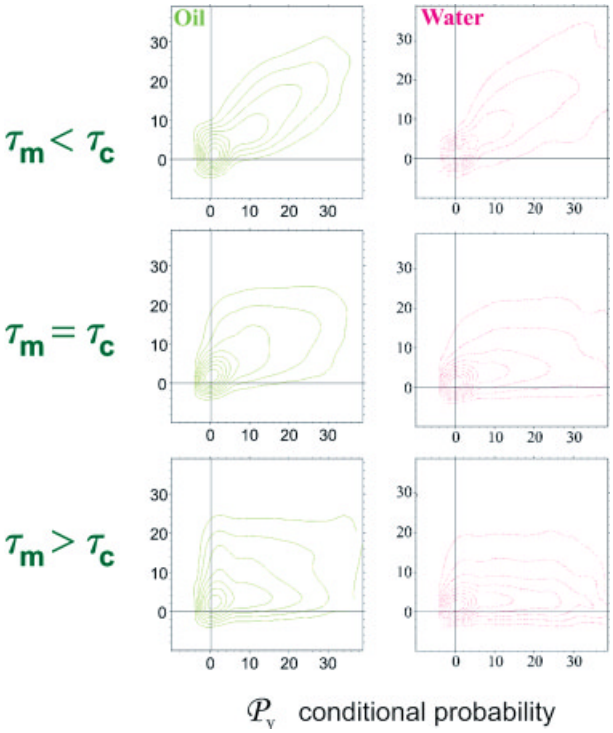

Figure 32. Example of a VEXSY experiment. In a setup similar to that shown in Figure 30B, water and silicone oil were simultaneously present in a random bed of monodisperse glass beads (400 mm diameter), and either of the two liquids was pumped with no net flow being applied to the other liquid. The average interstitial velocity was 9.5 mm/s (oil flow) and 11.0 mm/s (water flow). Data are shown for mixing times tm in relation to the characteristic correlation time of the system tc, indicating the time to travel one bead diameter. The function plotted is the conditional probability described in the text. The change from high correlation (alignment along the main diagonal) towards complete loss of correlation (horizontal contour lines) is evident with increasing tm. (reproduced from [54] with permission).

180° pulse gives rise to a further echo that is the starting point of the second displacement-encoding module (wave vector $\boldsymbol{q}_2 = \boldsymbol{k}_3 = -\boldsymbol{k}_4$). Only the final echo is acquired and contains the signal of spins that have experienced all four gradient pulses. The $\boldsymbol{q}$ vectors are usually applied in the same spatial direction so that the set of spins traveling at constant velocity will have identical R_1 and R_2 displacements, thus contributing to signal on the diagonal in the (R_1, R_2) space. By contrast, a migration of spins from one region of the displacement spectrum to another over the time τ_m will lead to off-diagonal contributions. The off-diagonal signal represents all spins that have changed their initial velocity, thus revealing complex motion including acceleration. Application of VEXSY, mainly to fluid-flow and particle-motion problems, has become increasingly important in recent years, and the interpretation of such correlations has been discussed by means of coordinate transformations [18, 43, 48] and 1D versions of the experiment [49 – 53]. An example of studies of two-phase flows in which the spectroscopic information of the echo is exploited is shown in Figure 32.

5.3. Velocity Maps

The methods that have been presented in Sections 5.1 and 5.2 generate pure statistical measures of motion parameters. However, it is frequently desirable to obtain a pictorial, and indeed also quantitative, representation of a flow field.

In order to determine the actual motion of a fluid with spatial resolution, and particularly for displacements smaller than the image resolution, the spins must be encoded with respect to position and velocity within a single pulse sequence; this is what can be called velocity imaging in

the "proper" sense. Before this can be achieved, it is vital to understand, and to avoid, the side-effects that flow has on an NMR image, i.e., the possibility of blurring or losing signal from certain volumes because the spins move out of the detection coil before they can contribute to the signal.

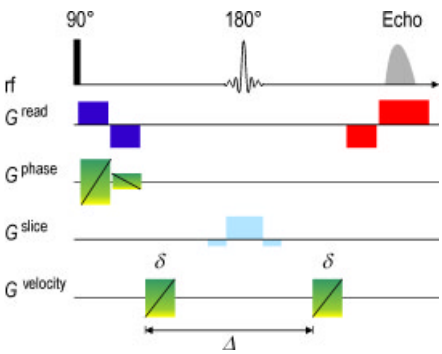

Figure 33. Typical motion-compensated, velocity-encoding imaging sequence. The basic scheme from Figure 8 is amended by additional compensating gradient pulses for read, phase, and slice directions. The velocity encoding pair of gradient pulses of duration δ and separation Δ is drawn on a separate line; in reality, it is applied along any of the three directions given for read, phase, and slice. The diagonal lines indicate that the magnitude of this gradient pulse pair is varied in a number of steps; the experiment, however, can also be executed with a single magnitude value by comparison to an equivalent sequence without the velocity-encoding gradients.

In fact, *blurring* — as would be observed in an optical photograph of a moving object — is an inadequate description of the NMR image artefacts arising due to motion, which can be much more complicated that just reducing the spatial resolution of an image. A plethora of ghosts, double and multiple images, unwanted signal attenuation, and so on, can arise, but once they become visible in the NMR image, they can be identified and taken care of. The best way to avoid artefacts is proper design of the pulse sequence that suppresses motion effects as completely as possible. Figure 33 shows a sequence which exploits the moment formalism explained in Section 5.2 to generate a correct velocity-encoded image. It is one of the simplest of such sequences and goes by the name "flow-compensated, velocity-encoding, imaging sequence". The meaning of this seemingly paradoxical name is the wish to introduce velocity encoding only in the direction in which it is desired (say, along the main axis of motion), but to suppress unwanted artefacts due to motion which might affect the image quality.

Relating to the moments discussion, the following requirements must be met in the sequence of Figure 33: $\boldsymbol{m}_0 \neq 0$, $\boldsymbol{m}_1 = 0$ in phase direction — the position encoding should be preserved, but it must be insensitive to motion — realized by a pair of rectangular gradients with equal duration and intensity ratio 3:1.

$\boldsymbol{m}_0 = 0$, $\boldsymbol{m}_1 = 0$ in slice direction — the gradient should excite a slice but leave no residual phase shift depending on either position or velocity — realized by a pair of negative lobes on both sides of the actual slice gradient with same total duration but one-third of the slice gradient strength.

$\boldsymbol{m}_0 = 0$, $\boldsymbol{m}_1 = 0$ in read (frequency-encoding) direction — these conditions can only be met for the center of the acquisition window. While $\boldsymbol{m}_0 \neq 0$ before and after this echo, so that position encoding is actually possible at all, it is unavoidable to have $\boldsymbol{m}_1 \neq 0$ during the same period because the evolution of the phase under the gradient is determined. By rendering $\boldsymbol{m}_1 = 0$ at the time of the echo, the influence of velocity on the signal is minimized, but in general the read direction will always be the one which is most severely affected by motion. Note that higher orders of motion can also be compensated if the gradient pulse trains fulfill conditions such as $\boldsymbol{m}_2 = 0$ and so on, at the cost of more complicated and longer encoding schemes (see [42] for a useful overview).

In order to actually generate a velocity image, in addition to the compensated phase, slice, and read gradients, a gradient pair is introduced which has the purpose of encoding velocity. It can be applied in any of the three directions, and at any time within the sequence. However, practical experience — or a thorough computation of the combined effect of all applied gradients, as well as the often unavoidable background gradients and susceptibility effects — suggests short and symmetric sequences, with the flow-encoding direction coincident with the phase-encoding direction if image *and* flow information is to be encoded along the same direction, and read gradients along the direction of slowest motion. The best solution for a particular application cannot be given in general and will

depend on many factors such as field strength, available hardware, sample shape and material, magnitude of velocities, and many more.

Figure 33 contains all that is necessary to generate a three-dimensional image with a three-dimensional velocity field overlaid on it. The actual way in which this image is generated depends somewhat on the expected result, primarily on the range of velocities.

Velocity information is brought into the signal by the bipolar encoding gradient via a phase shift according to Equation 19. If one were satisfied with the measure of an average velocity in a pixel (or there is reason to assume that velocities within one pixel do not vary much), one can use this phase shift directly and translate it into velocities with the proportionality $\phi(\boldsymbol{v}) = 2\pi \boldsymbol{q} \cdot \boldsymbol{v} \Delta$. (Note that instead of displacements, the result is expressed in terms of velocities, which is fully equivalent, provided that velocities within the pixel do not change significantly during the encoding time Δ.) However, since the image-reconstruction process itself also relies on phase information, the bipolar velocity-encoding gradient just adds an extra contribution that cannot easily be separated from the phase shifts resulting from the imaging gradients. The solution is to acquire two images: one with and one without ($\boldsymbol{q} = 0$) the velocity-encoding gradient. Both images are processed, and the phases in each pixel are computed; because the signal is complex, the phase angle is given by $\varphi = \tan^{-1}[Im(S)/Re(S)]$, where *Im* and *Re* denote the imaginary and the real part of the signal, respectively. One can then subtract the phases of each pixel, and the difference directly renders the velocity in that pixel. In a real experiment, imperfect field and gradient distributions and local susceptibility variations can, and usually will, give rise to uncontrollable shifts of the reference phase, which are eliminated by this subtraction procedure as well.

The result of such a procedure contains the information of the spin density of the object and the average velocity per pixel in one direction. One can simply repeat the experiment with the vector $\boldsymbol{q}$ aligned in the other two orthogonal directions and obtain a set of data that can be used to reconstruct a vector plot, where each vector contains all velocity components in each pixel. Because of the reference measurement, four individual experiments are needed to achieve this. However, the flow field of interest is often described sufficiently by certain components of the velocity, or by two-dimensional images corresponding to selected slices of the object. Two examples of vector plots in quite different situations — internal circulation in water drops and gas flow around an obstacle — are shown in Figure 34.

Whenever a finite *velocity distribution* within a pixel is suspected, the propagator can be obtained by using the same pulse sequence, but varying the magnitude of the velocity-encoding wave vector in full analogy to the measurement of the pure, spatially unresolved propagator described in Section 5.2. This case is shown in Figure 33, where the diagonal lines in the flow-encoding gradient symbolize repeated application with different values. As in all propagator and velocity encoding experiments, the range of gradients must be considered. In complete analogy to encoding of position (Eq. 13), the range of velocities covered, also called field of flow (FOF), is given by:

$$\mathrm{FOF} = \frac{n-1}{2} \frac{2\pi}{\gamma g_{\max} \delta \Delta} \tag{29}$$

Like in imaging, aliasing can occur if velocities outside the FOF exist. For measurements of average velocities, the same condition holds if the total phase shift due to velocity exceeds 2π. For velocity fields of simple structure, this problem can be accounted for by so-called unwrapping algorithms.

5.4. Time-of-Flight Methods

For high velocities and displacements exceeding one pixel, or in general for situations where the visualization of flow and obstructions is more relevant than a quantitative evaluation, a number of quite different experiments have been designed that are based on a time lag between excitation and image acquisition, commonly known as time-of-flight methods.

Several methods exploit the transport of spins or dopants into or out of the image volume to derive a measure of velocities. For instance, if an image is recorded in a particular slice of the excitation volume of the resonator by using a slice-selective rf pulse, fluid leaving this slice during the pulse sequence before acquisition will lead

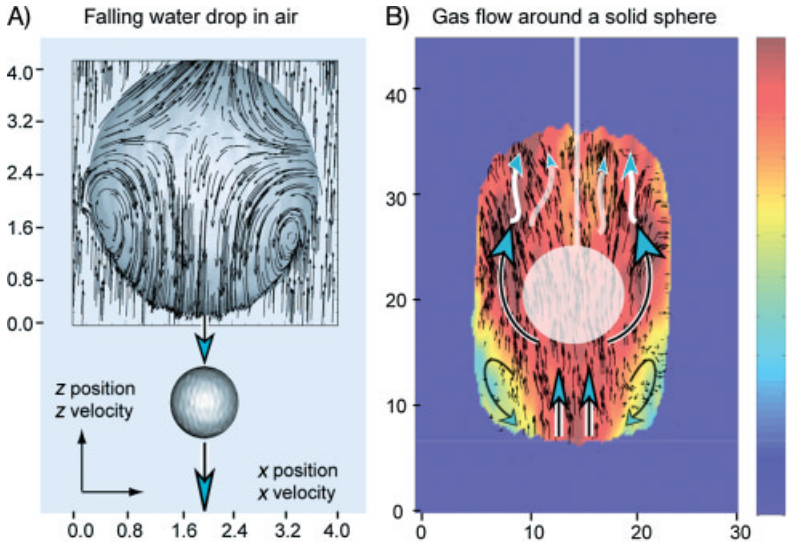

Figure 34. A) Vector plot of velocities inside a water drop falling in air [55]. Each individual encoding step was applied to a different drop, with drops triggering the acquisition time by passing a light gate. Care was taken to ensure that all drops were identical in size and motion, thereby allowing the reconstruction of an average flow pattern. The vortex structure of the internal motion, driven by air friction, is clearly visible. The drop was falling at 2 m/s, and maximum internal velocities were on the order of ± 0.2 m/s. B) Vector plot of gas flow around a 10 mm diameter solid sphere. For this experiment, the signal from hyperpolarized ^{129}Xe nuclei in a gas mixture was used; the maximum gas velocity was 180 mm/s.

to a loss of signal intensity, as fresh fluid that enters into this slice has not been encoded by the pulse sequence. This approach is often used in angiography, the medical discipline of monitoring blood transport in vessels, and is known as artery spin labeling (ALS). Another possibility is the addition of a tracer substance that has an effect on NMR parameters, for instance, a paramagnetic complex that reduces relaxation times when dissolved in water. The spatial distribution of relaxation times determined in a T_1 or T_2 parameter image allows the reconstruction of the transport properties of the tracer in a series of images. The technique is thus favorable for slow transport processes on timescales of seconds to hours, or for qualitative visualizations where the main objective is to detect the presence of the tracers in part of the sample.

A related strategy is introduced by so-called tagging or, more generally speaking, time-of-flight experiments. The tagging technique is a different approach to determining velocities in a moving object. It is often used for visualization, but can also be extended to provide quantitative information about flow fields. The principle consists of "tagging", or labeling, a certain region of the sample by manipulating the magnetization inside this volume and acquiring the image after some delay during which the labeled volume is deformed. For flow systems with time-invariant flow patterns, movies can be reconstructed from a series of such images with different delay periods. A simple realization is the tagging of a single slice, which is achieved by applying a soft pulse in the presence of a pulsed field gradient, exactly as described in Section 2.3 for slice selection. This slice can be "dark", that is, the spins in the slice are saturated by the rf pulse and the subsequent imaging sequence generates an image of the total volume minus the saturated spins, which show up as a dark stripe. Figure 35 shows one possibility of what such a pulse sequence can look like, and snapshots for flow of a fluid through a fixed bed of catalyst containing pellets are presented as an example for this method. By applying the slice-selective pulse at a different position within the sequence, or simply by subtracting the image from a reference spin-density image acquired without a tagging element, images can be obtained which highlight the moving volume in front of a dark background, as has been done in this example.

A tagging experiment often makes complicated motion more readily accessible for the ex-

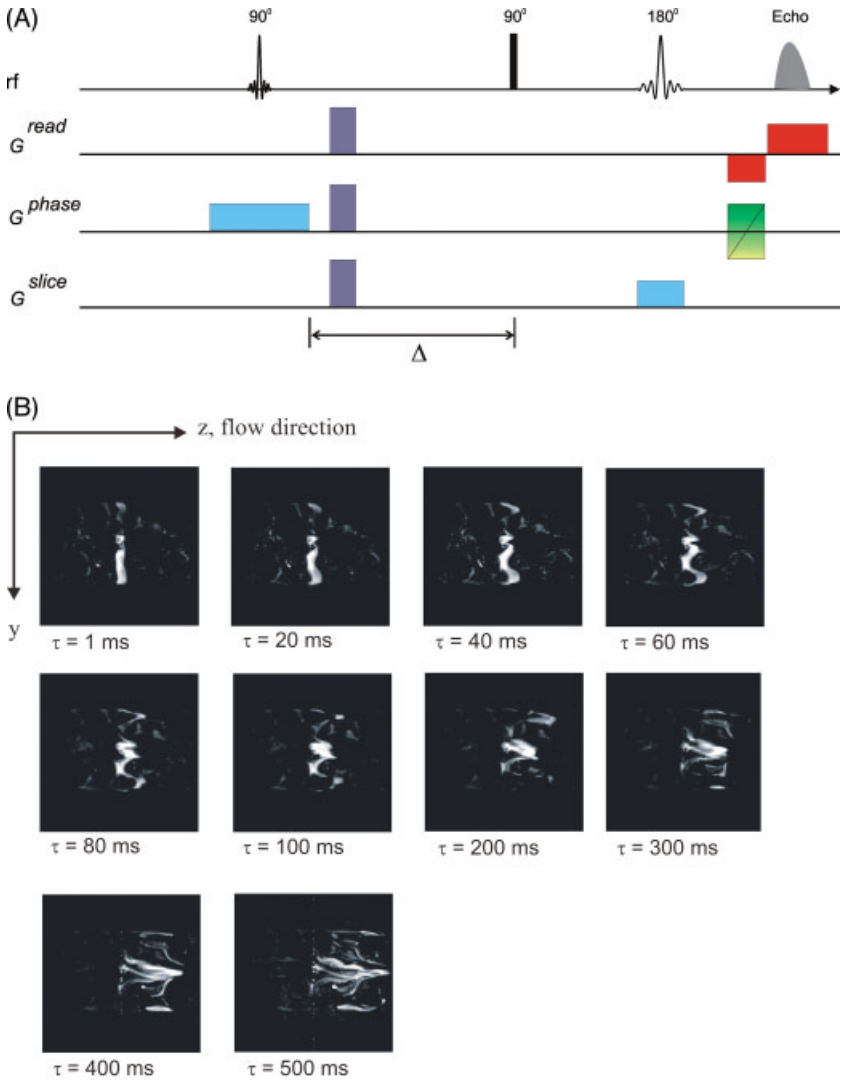

Figure 35. A) Single-slice tagging sequence. The 90° soft pulse in the presence of a magnetic field gradient excites the magnetization inside a slice and transfers it to the transverse plane. The subsequent so-called spoiler gradients destroy any coherence of this magnetization so that it will not contribute to the signal later on in the sequence (the spoiler gradients are drawn in three orthogonal directions, but a single direction is often sufficient). The following imaging sequence acquires a signal of the whole object with the exception of the spins inside the slice that was excited in the beginning and that has been deformed due to motion during the waiting time Δ; a single black stripe inside an otherwise undisturbed image is observed. B) Tagging of water flow in a bed packed with 7 mm diameter beads. The width of the grid line is 2 mm. The evolution time varied τ varied between 0 and 500 ms. The volume flow rate is 2.93 mm^3/s, and the spatial resolution 195 × 390 mm. Each image is the result of a tagging experiment with a given evolution time which has been subtracted from a reference frame to highlight the deformed slice. (Reproduced from [56] with permission).

perimenter, but is limited to displacements exceeding the image resolution, or when fast image acquisition of the flow phenomenon is needed. The ultimate time limit, for phase encoding and for tagging, is the longitudinal relaxation time of typically a few seconds in liquids; if it is much exceeded, the information encoded into the spin system is lost, or — for tagging — the contrast between tagged spins and the environment has disappeared. For an additional overview of motion measurements in industrial applications, see [24, 56 – 59].

6. Hardware and Materials

6.1. Imaging and Micro-Imaging Hardware

It is not the purpose of this article to provide a complete list of NMR imaging equipment and manufacturers. However, some general remarks can be given that allow the reader to assess the feasibility of studying a particular problem.

Some main features that need to be discussed in the context of NMR imaging are sample dimensions; magnetic field strengths; field homogeneities; rf and gradient capabilities; and data acquisition, storage, and processing properties.

The first concern of the user is certainly the shape and size of the sample. Because of the need to generate homogeneous magnetic fields for the majority of studies, the object of interest usually has to be surrounded by the magnet. Solenoid coils provide a maximum of field homogeneity, but the final optimization is achieved by so-called shim systems, resistive coils wound in particular patterns to generate magnetic field corrections up to fourth or fifth order when the field dependence is expressed in a Taylor series expansion relative to the spatial coordinates. The larger the sample volume, the more correction coils are required to achieve a "flat" field distribution, but the extreme degree of homogeneity used in spectroscopy for samples of only some millimeters in diameter cannot be obtained in an imaging environment.

Imaging systems can often be attributed to one of two standard basic setups. On the one hand, whole-body systems employed in medical diagnostics for human patients have the largest available volumes; on the other hand, standard superconducting magnets produced for spectroscopy are frequently used for microimaging purposes, leaving diameters of a few centimeters for the sample. In between, dedicated magnets of typically 10 – 20 cm accessible diameter are also available that serve the purpose of animal studies.

Spatial resolution usually goes along with sample, or rather resonator, dimensions. Considering the experimental time required for repeated phase encoding, a rule of thumb says that the resolution is about 1/100 of the total dimension. That, in turn, means that ever stronger gradients are required if one wants to reach this resolution limit in small samples. On the other hand, because gradient coils also need to enclose the whole sample, the total power consumption grows drastically towards large-object imaging. For example, typical maximum gradient strengths for medical scanners are 30 – 50 mT/m at currents of several hundred amperes, while a standard microimaging system works at 1 – 2 T/m with a current of 50 A. For true microscopic resolution, several tens of teslas per meter are applied in gradient coils of much smaller dimensions. Total signal intensity and heating effects are some of the limits at the small scale. Note that it is often not the maximum gradient strength itself, but the capability of switching gradient currents in a fast and reproducible way, that determines the feasibility of fast and multiscan experiments.

The magnetic field strength is the other main characteristic of the NMR apparatus. In general, high fields are desired because of their direct influence on the signal-to-noise ratio (exceptions are samples with high internal susceptibility gradients, where low fields can possess advantages), but they generally come at higher costs. Again, one frequently relies on standard magnetic fields that are commercially available — between 1 and 4 T for medical imaging (a range of systems working at lower field have been developed recently), and between 4.7 T and 18 T for standard laboratory magnets, corresponding to ^{1}H Larmor frequencies between 200 and 750 MHz.

One further topic of NMR experiments is the capability of rf generation as well as the data acquisition parameters. Since the requirements for multidimensional NMR spectroscopy (multichannel pulse generation; precise reproducibility of frequencies, phases, amplitudes, and times; digital control over pulse shapes and bandwidth) are usually far more demanding, MRI equipment has benefited considerably from developments in the spectroscopy sector. Likewise, fast data acquisition and on-line image reconstruction have become routine aspects due to the need to optimize clinical scan parameters by immediate verification of the imaging quality.

In recent years, a growing number of low-field devices based on electromagnets or permanent magnets have been developed that meet two common goals: to reduce cost and thus make available imaging systems for the nonspecial-

ist and for routine applications, and to bring the sensor to the object in order to increase flexibility and provide in-line or on-line capabilities in production processes. Of course, this approach comes at a price (e.g., reduced sensitivity and field homogeneity), but imaging, particularly of ^{1}H-containing substances, is still feasible within timescales comparable to high-field devices because, once the signal-to-noise ratio remains above a certain level, the experimental time is limited only by the spatial resolution but not by the number of accumulations. The decrease of T_1 with lower field strength that is observed for many types of samples (except free liquids) is actually of advantage for performing faster experiments. Several types of these systems are currently becoming commercially available.

Apart from classical imaging systems, i.e., magnets with relatively homogeneous fields with defined triaxial gradients [60, 61], a number of alternative approaches have been developed that allow the acquisition of pointwise information or the scanning of objects with 1D or 2D resolution. These can be of considerable advantage whenever the full 3D contrast information is not required, for instance, in monitoring of local relaxation parameters or depth-dependent profiling. The main advantage of these scanners is their high flexibility in obtaining NMR information on surfaces and extended objects in arbitrary places, making use of the portability of the complete system consisting of the sensor and the amplifier/acquisition/analysis unit [62, 63].

6.2. Imaging X Nuclei

Imaging of protons represents the majority of all MRI applications. This is mostly due to the high sensitivity of the ^{1}H nucleus, which has the largest gyromagnetic ratio among all stable isotopes, but also because of its ubiquity in most systems of interest, both medical and nonmedical. However, other nuclei are of substantial interest, mostly because their visualization is more specific, or because these are more sensitive as tracers for their local environment. For example, ^{1}H has one of the smallest ranges of chemical shifts, and ^{13}C chemical shift imaging is much more suitable for generating distribution maps of individual substances. Imaging, with its requirements of extended measurement time, depends on the abundance and receptivity of the NMR-active nuclei much more than conventional spectroscopy. For instance, quadrupolar nuclei with intrinsically large line widths and short T_2 often cannot be used in time-consuming imaging sequences, for example, if long echo times are involved. The discussion in → Nuclear Magnetic Resonance and Electron Spin Resonance Spectroscopy retains its validity, but the number of technically useful nuclei in imaging is smaller. Table 1 summarizes the most important isotopes for NMR imaging.

Despite the low sensitivity to many X nuclei, the principle of NMR imaging remains unchanged, so that discussion of this topic is intentionally kept short. However, it must be stressed that the disadvantage of weak signal is often compensated by the gain in information, and a number of tools are available to increase the sensitivity towards X nuclei substantially. One possibility is the hyperpolarization technique, which mostly applies to gases (see Section 6.3). Many more possibilities are obtained by borrowing strategies known from NMR spectroscopy that are used to transfer polarization from abundant to rare nuclei. For instance, polarization is transferred from ^{1}H to X nuclei, from X to ^{1}H (where the detection sensitivity is higher), or in a cyclic way from ^{1}H to X and back to ^{1}H. Techniques known as DEPT, [65] INEPT, and HMQC not only increase the signal intensity of, for example, ^{13}C nuclei, but also enhance the selectivity, because polarization transfer is achieved only for proton-bound carbon atoms. In addition, less specific transfer sequences exist, all having in common a carefully timed and balanced series of rf pulses on at least two separate frequency channels, this being the minimum hardware requirement for this type of signal enhancement of X nuclei.

Other approaches that currently are gaining increasing importance are based on the transfer of polarization from a suitable molecule or a radical to the nuclei of interest, where the polarization lifetime and efficient transfer are the main problems to be overcome. However, it has even become possible to apply para-hydrogen-induced polarization (PHIP) [66] to in vivo angiography by transferring the PHIP polarization to ^{13}C of a tracer molecule which was subsequently injected into the bloodstream

Table 1. The most important nuclei used in NMR imaging. Larmor frequencies are given for a typical magnetic field strength of modern spectrometers. The molar receptivity relative to ^{1}H assumes equal numbers of nuclei; it is multiplied by the natural abundance to obtain the corresponding receptivity; the numbers assume thermal equilibrium. Isotopic enrichment is routinely applied for studies of ^{2}H, ^{13}C, and ^{17}O, but is also common for ^{3}He and ^{129}Xe. The latter two nuclei are marked in gray to indicate that they are candidates for imaging experiments in the hyperpolarized state. Under these circumstances, their signal strength can be several orders larger than their thermal receptivities. (Values taken from [64])

Isotope	Spin	Natural abundance, %	Larmor frequency at 11.7 T, MHz	Molar receptivity	Receptivity at natural abundance
^{1}H	1/2	99.988	500	1	1
^{2}H	1	0.012	76.75	9.6×10^{-3}	1.1×10^{-6}
^{3}He	1/2	1.4×10^{-4}	380.90	0.44	6.2×10^{-5}
^{13}C	1/2	1.07	125.73	1.6×10^{-2}	1.7×10^{-4}
^{17}O	5/2	0.038	67.78	2.9×10^{-2}	1.1×10^{-5}
^{19}F	1/2	100	470.47	0.83	0.83
^{23}Na	3/2	100	132.26	9.3×10^{-2}	9.3×10^{-2}
^{31}P	1/2	100	202.40	6.7×10^{-2}	6.7×10^{-2}
^{129}Xe	1/2	26.4	139.05	2.2×10^{-2}	5.7×10^{-3}

[67]. Dynamic nuclear polarization (DNP), that is, the transfer of electron magnetization to nuclei via stable radicals introduced into the sample, is developing into a routine method as problems with fast sample heating and transport have been overcome. It requires transfer of polarization from unpaired electrons to nuclei in the solid (usually low-temperature) phase and subsequent heating preserving the nuclear states sufficiently long for performing single-scan experiments with dramatic increase in sensitivity [68].

6.3. Liquids, Solids, and Gases

Most imaging sequences — including those presented in this contribution — are designed for liquids or soft matter with their favorable properties such as relatively long T_2 and moderate T_1 relaxation times. However, solids are also of substantial interest, for example, in materials research. Dedicated methods such as single-point imaging are often the only way to enable imaging of solids. For samples with slightly longer transverse relaxation times, echo methods or frequency encoding and backprojection techniques represent another possibility.

The situation is somewhat different for gases. Depending on the mechanism of relaxation, either very short or very long relaxation times are observed. In monatomic noble gases, relaxation times can be on the order of minutes to hours and depend strongly on the presence of other nuclei or surfaces that serve as relaxation sinks. While this appears particularly suitable for the implementation of multi-echo sequences, the large diffusivity of gases severely restricts the use of pulsed field gradients because they lead to an irreversible loss of coherence. Therefore, again, position encoding must be performed on a short timescale, and to overcome the long longitudinal relaxation times, small flip angles such as in FLASH, are often required. The extraordinary advantage of noble gases due to their low relaxivity, however, is the possibility to polarize them almost to the theoretical maximum in a process called hyperpolarization [69]; ^{129}Xe and ^{3}He have been used extensively for these purposes, and polarization equipment has meanwhile become available commercially. For combining hyperpolarization techniques with measurements of transport, see Figure 34b.

Polyatomic gases — most importantly hydrocarbons or fluorinated gases, which are of particular importance due to their high density and the natural contrast in a fluorine-free environment— have a different relaxation mechanism, and values of T_1 and T_2 are in the range of only a few milliseconds or less. With the opposite argument to monatomic gases, this leads to the considerable advantage of rapid repetition times so that multi-echo sequences are replaced by a series of individual phase- and frequency-encoded sequences. This property compensates, to a certain degree, the disadvantages brought about by a spin density which is orders of magnitude lower than that of liquids.

7. References

1. A. Abragam: *The Principles of Nuclear Magnetism*, Clarendon Press, Oxford 1961.

2. C. P. Slichter: *Principles of Magnetic Resonance*, 3rd ed., Springer Verlag, Berlin 1996.
3. R. Freeman: *A Handbook of Nuclear Magnetic Resonance*, 2nd ed., Addison Wesley Longman, Essex 1997.
4. R. Freeman: *Magnetic Resonance for Chemistry and Medicine*, Oxford University Press, Oxford 2003.
5. E. Fukushima, S. B. W. Roeder: *Experimental Pulse NMR — a Nuts and Bolts Approach*, 10th ed., Perseus Publishing, Jackson, TN 1986.
6. R. Kimmich: *NMR. Tomography, Diffusometry, Relaxometry*, Springer, Berlin/Heidelberg 1997.
7. R. K. Harris: *Nuclear Magnetic Resonance Spectroscopy*, Longman Scientific & Technical, Harlow 1986.
8. R. R. Ernst, G. Bodenhausen, A. Wokaun: *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, Clarendon Press, Oxford 1990.
9. H. Günther: *NMR Spectroscopy: Basic Principles, Concepts and Applications in Chemistry*, Wiley, New York 1995.
10. R. Freeman: *Spin Choreography, Basic Steps in High Resolution NMR*, Oxford University Press, Oxford 1998.
11. R. S. Macomber: *A Complete Introduction to Modern NMR Spectroscopy*, Wiley Interscience, New York 1998.
12. S. Braun, S. Berger: *200 and More Basic NMR Experiments*, Wiley-VCH, Weinheim 2004.
13. H. Friebolin: *Basic One- and Two-Dimensional NMR Spectroscopy*, 4th ed., Wiley-VCH, Weinheim 2005.
14. D. M. Grant, R. K. Harris (eds.): *Encyclopedia of Nuclear Magnetic Resonance*, Wiley, Chichester 1996.
15. G. Webb (ed.): *Handbook of Modern Magnetic Resonance*, Springer, Amsterdam 2006.
16. I. Serša, S. Macura, "Excitation of Arbitrary Shapes by Gradient Optimized Random Walk in Discrete k-Space", *Magn. Reson. Med.* **37** (1997) 920.
17. Paul T. Callaghan: *Principles of Nuclear Magnetic Resonance Microscopy*, Clarendon Press, Oxford 1995.
18. B. Blümich: *NMR Imaging of Materials*, Clarendon Press, Oxford 2000.
19. D. W. McRobbie, E. A. Moore, M. J. Graves, M. R. Prince: *MRI — From Picture to Proton*, Cambridge University Press, Cambridge 2003.
20. M. T. Vlaardingerbroek, J. A. Den Boer: *Magnetic Resonance Imaging: Theory and Practice*, 2nd ed., Springer, Berlin/Heidelberg 2004.
21. E. M. Haacke, R. W. Brown, M. R. Thompson, R. Venkatesan: *Magnetic Resonance Imaging: Physical Principles and Sequence Design*, Wiley-Liss, New York 1999.
22. M. A. Bernstein, K. F. King, X. J. Zhou: *Handbook of MRI Pulse Sequences*, Academic Press, San Diego 2004.
23. P. Blümler, B. Blümich, R. E. Botto, E. Fukushima (eds.): *Spatially Resolved Magnetic Resonance: Methods, Materials, Medicine, Biology, Rheology, Geology, Ecology, Hardware*, VCH, Weinheim 1998.
24. S. Stapf, S. Han (eds.): *NMR Imaging in Chemical Engineering*, Wiley-VCH, Weinheim 2006.
25. J. Kowalewski, L. Maler: *Nuclear Spin Relaxation in Liquids*, Taylor & Francis, London 2006.
26. I. Bertini, C. Luchinat, G. Parigi: *Solution NMR of Paramagnetic Molecules*, Elsevier, Amsterdam 2006.
27. A. E. Merbach, E. Toth: *The Chemistry of Contrast Agents in Medical Magnetic Resonance Imaging*, Wiley, New York 2001.
28. A. T. Watson, C. T. P. Chang, "Characterizing Porous Media with NMR Methods", *Progr. Nucl. Magn. Reson. Spectrosc.* **31** (1997) 343.
29. R. Kimmich, D. Hoepfel, "Volume-Selective Multipulse Spin-Echo Spectroscopy", *J. Magn. Reson.* **72** (1987) 379.
30. L. F. Gladden, L. D. Anadon, M. H. M. Lim, A. J. Sederman, "Two-Phase Flow in Trickle-Bed Reactors" in S. Stapf, S. Han (eds.): *NMR Imaging in Chemical Engineering*, Wiley-VCH, Weinheim 2006.
31. A. Haase, J. Frahm, D. Matthei, W. Hänicke, K. D. Merboldt: "FLASH Imaging: Rapid NMR Imaging Using Low Flip Angles", *J. Magn. Reson.* **67** (1986) 258.
32. B. J. Balcom, R. P. MacGregor, S. D. Beyea, D. P. Green, R. L. Armstrong, T. W. Bremner, "Single-Point Ramped Imaging with T Enhancement (SPRITE)", *J. Magn. Reson.* A **123** (1996) 131.
33. J. J. Young, T. W. Bremner, M. D. A. Thomas, B. J. Balcom: "Pure Phase Encode Magnetic Resonance Imaging of Concrete Building Materials" in S. Stapf, S. Han (eds.): *NMR Imaging in Chemical Engineering*, Wiley-VCH, Weinheim 2006.
34. A. A. Khrapitchev, B. Newling, B. J. Balcom, "Sectoral Sampling in Centric-Scan SPRITE

Magnetic Resonance Imaging", *J. Magn. Reson.* **178** (2006) 288.
35. J. Hennig, A. Nauerth, H. Freidburg, "RARE Imaging: A Fast Imaging Method for Clinical MR", *Magn. Reson. Med.* **3** (1986) 823.
36. P. Mansfield, P. G. Morris, *Adv. Magn. Res*. Suppl. 2, Academic Press, New York 1982.
37. S. Stapf, S. Han, C. Heine, B. Blümich, "Spatio-Temporal Correlations in Transport Processes Determined by Multi-PFG Experiments", *Concepts Magn. Reson.* **14** (2002) 172.
38. J. Kärger, W. Heink, "The Propagator Representation of Molecular-Transport in Microporous Crystallites", *J. Magn. Reson.* **51** (1983) 1.
39. W. S. Price, "Pulsed-Field Gradient Nuclear Magnetic Resonance as a Tool for Studying Translational Diffusion: Part 1. Basic Theory", *Concepts Magn. Reson.* **9** (1997) 299.
40. W. S. Price, "Pulsed-Field Gradient Nuclear Magnetic Resonance as a Tool for Studying Translational Diffusion: Part 2 Experimental Aspects", *Concepts Magn. Reson.* **10** (1998) 197.
41. E. Fukushima, "Nuclear Magnetic Resonance as a Tool to Study Flow", *Annu. Rev. Fluid Mech.* **31** (1999) 95.
42. J. M. Pope, S. Yao, "Quantitative NMR Imaging of Flow", *Concepts Magn. Reson.* **5** (1993) 281.
43. S. Han, S. Stapf, B. Blümich, "Two-Dimensional PFG NMR for Encoding Correlations of Position, Velocity, and Acceleration in Fluid Transport" *J. Magn. Reson.* **146** (2000) 169.
44. A. J. Sederman, M. D. Mantle, C. Buckley, L. F. Gladden, "MRI Technique for Measurement of Velocity Vectors, Acceleration, and Autocorrelation Functions in Turbulent Flow", *J. Magn. Reson.* **166** (2004) 182.
45. B. Buhai, A. Hakimov, I. Ardelean, R. Kimmich, "NMR Acceleration Mapping in Percolation Model Objects", *J. Magn. Reson.* **168** (2004) 175.
46. S. Stapf, K. J. Packer, R. G. Graham, J. F. Thovert, P. M. Adler, "Spatial Correlations and Dispersion for Fluid Transport through Packed Glass Beads", *Phys. Rev. E* **58** (1998) 6206.
47. P. T. Callaghan, B. Manz, "Velocity Exchange Spectroscopy", *J. Magn. Reson. A* **106** (1994) 260.
48. S. Han, B. Blümich: "2-Dimensional Representation of Position, Velocity and Acceleration by PFG-NMR", *Appl. Magn. Reson.* **18** (2000) 101.
49. P. P. Mitra: "Multiple Wave-Vector Extensions of the NMR Pulsed-Field Gradient Spin-Echo Diffusion Measurement", *Phys. Rev. B* **51** (1995) 15 074.
50. Y. Cheng, D. G. Cory: "Multiple Scattering by NMR", *J. Am. Chem. Soc.* **121** (1999) 7935.
51. P. T. Callaghan, S. L. Codd, J. D. Seymour: "Spatial Coherence Phenomena Arising from Translational Spin Motion in Gradient Spin Echo Experiments", *Concepts Magn. Reson.* **11** (1999) 181.
52. S. L. Codd, B. Manz, J. D. Seymour, P. T. Callaghan: "Taylor Dispersion and Molecular Displacements in Poiseuille Flow", *Phys. Rev. E* **60** (1999) R3491.
53. A. Caprihan, J. D. Seymour: "Correlation Time and Diffusion Coefficient Imaging: Application to a Granular Flow System", *J. Magn. Reson.* **144** (2000) 96.
54. A. A. Khrapitchev, S. Han, S. Stapf, B. Blümich: "Spectrally Resolved Velocity Exchange Spectroscopy of Two-Phase Flow", *J. Magn. Reson.* **159** (2002) 36.
55. S. Han, S. Stapf, B. Blümich, "NMR Imaging of Falling Water Drops", *Phys. Rev. Lett.* **87** (2001) 144501.
56. X. Ren, S. Stapf, B. Blümich: "Multiscale Approach to Catalyst Design" in S. Stapf, S. Han (eds.): *NMR Imaging in Chemical Engineering*, Wiley-VCH, Weinheim 2006.
57. L. F. Gladden, P. Alexander: "Applications of Nuclear Magnetic Resonance Imaging in Process Engineering", *Meas. Sci. Technol.* **7** (1996) 423.
58. P. T. Callaghan: "Rheo-NMR: Nuclear Magnetic Resonance and the Rheology of Complex Fluids", *Rep. Prog. Phys.* **62** (1999) 599.
59. L. F. Gladden, "Recent Advances in MRI Studies of Chemical Reactors: Ultrafast Imaging of Multiphase Flows", *Top. Catal.* **24** (2003) 19.
60. K. Kose et al.: "Development of a Compact MRI System for Trabecular Bone Volume Fraction Measurements", *Magn. Reson. Med.* **52** (2004) 440.
61. B. Goodson: "Mobilizing Magnetic Resonance", *Phys. World* **5** (2006) 28.
62. J. Perlo, F. Casanova, B. Blümich: "Profiles with Microscopic Resolution by Single-Sided NMR", *J. Magn. Reson.* **176** (2005) 64.
63. J. Mitchell, P. Blümler, P. J. McDonald: "Spatially Resolved Nuclear Magnetic

Resonance Studies of Planar Samples", *Progr. Nucl. Magn. Reson. Spect.* **48** (2006) 161.
64. Bruker Biospin Almanac, 2003.
65. B. S. Akpa, M. D. Mantle, A. J. Sederman, L. F. Gladden: "In situ C-13 DEPT-MRI as a Tool to Spatially Resolve Chemical Conversion and Selectivity of a Heterogeneous Catalytic Reaction Occurring in a Fixed-Bed Reactor", *Chem. Commun.* **21** (2005) 2741.
66. J. Natterer, J. Bargon: "Parahydrogen Induced Polarization", *Progr. Nucl. Magn. Reson. Spectr.* **34** (1999) 71.
67. K. Golman et al.: "Parahydrogen-Induced Polarization in Imaging: Subsecond ^{13}C Angiography", *Magn. Reson. Med.* **46** (2001) 1.
68. J. H. Ardenkjaer-Larsen et al.: "Increase in Signal-to-Noise Ratio of $>$ 10,000 Times in Liquid-State NMR", *Proc. Nat. Acad. Sci. USA* **100** (2003) 10158.
69. T. G. Walker, W. Happer: "Spin-Exchange Optical Pumping of Noble-Gas Nuclei", *Rev. Mod. Phys.* **69** (1997) 629.

Radial-Flow Packed-Bed Reactors

JEFF C. H. LI, Jacobs Engineering, Golden, Colorado, USA

1. **Introduction** 1
2. **Applications of Radial-Flow Reactors** . 2
3. **Design of Commercial Radial-Flow Reactors** 6
4. **Flow Distribution** 6
5. **Catalyst Basket Design** 8
6. **Catalyst Settling and Flow Bypassing** . 9
7. **Acknowledgments** 12
8. **References** 12

The radial-flow packed-bed reactor is a design with very small pressure drop and most suitable for large-capacity plants. It has replaced the axial-flow reactor for several important chemical processes such as ammonia synthesis, ethylbenzene dehydrogenation, and catalytic reforming. The design of a radial-flow reactor is more complicated than that of an axial-flow reactor; it must deal with several unique concerns: flow distribution along the length of the reactor, thermal reaction in the void space, catalyst settling, and expansion and contraction of the catalyst basket. The successful operation of a commercial radial-flow reactor lies mainly in the details of the mechanical design.

1. Introduction

Radial-flow packed-bed reactors are used to carry out several important, large-scale catalytic chemical reactions in the petroleum and chemical industries, such as ammonia synthesis, methanol synthesis, catalytic reforming, and dehydrogenation of ethylbenzene to styrene.

Figure 1 shows how radial-flow reactors differ from the more commonly used axial-flow reactors. In an axial-flow reactor, the feed enters at one end of the reactor, flows across the catalyst bed in the direction along the axis of the reactor, and exits from the other end. In a radial-flow reactor, the feed is distributed along the length of the reactor, and flows in the radial direction across the catalyst bed contained between two perforated concentric screens (the screen assembly that contains the catalyst is commonly called a basket). The reactant may flow from the center toward the outer annulus of the reactor or from the outer annulus toward the center of the reactor. In a Z-shaped radial-flow reactor, the feed and the effluent flow in the same direction in the center pipe and the annulus; in a Π-shaped reactor, the feed and the effluent move in the opposite directions. As discussed in Chapter 4, different flow arrangements—inward versus outward flow and Z-shaped versus Π-shaped—have been used for commercial reactor designs.

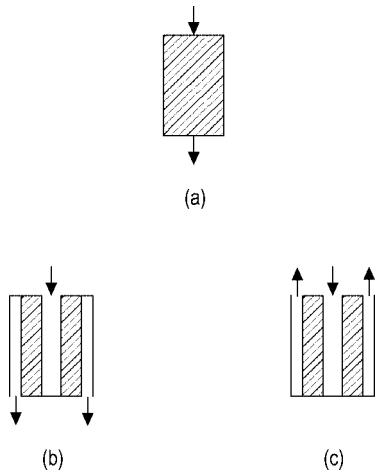

Figure 1. Axial- versus radial-flow reactors. a) Axial-flow reactor; b) Z-shaped radial-flow reactor; c) ?Π-shaped radial-flow reactor

Radial-flow reactors were developed to replace axial-flow reactors for large-scale chemical plants in the late 1960s and early 1970s [1 – 7]. As the plant capacity increases, the diameter of an axial-flow reactor must be increased in order to maintain the same pressure drop across the catalyst bed. With a larger diame-

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.l22_l01

ter, the reactor shell becomes thicker, heavier, and more expensive to fabricate. Besides, the diameter of the reactor is limited by the capabilities of fabrication and transportation facilities. Above some plant size, single axial-flow reactors become impractically large, and multiple reactors in parallel become necessary. In contrast, a radial-flow reactor can be scaled up by increasing the length instead of the diameter. Furthermore, since in a radial-flow reactor the reactant is fed along the length of the reactor, a larger cross-sectional area is available for the flow and the catalyst bed is shallower; as a result, the pressure drop across a radial-flow catalyst bed is much less than that across an axial-flow catalyst bed.

An example of the pressure-drop advantage of a radial-flow reactor over an axial-flow reactor is shown in Table 1. For a 1500 t/d ammonia plant, the pressure drop across the catalyst beds of a radial-flow reactor is only 0.2 kg/cm^2 versus 12 kg/cm^2 across those of an axial-flow reactor, and the power consumption of the synthesis loop with a radial-flow reactor is 64 % of that with an axial-flow reactor. Many old ammonia plants were retrofitted with radial-flow catalyst beds to increase plant capacity and reduce energy consumption [9 – 11]. Because the pressure drop across the radial-flow reactor is very small, this change also allows a reduction in catalyst particle size from 6–10 mm to 1.5–3.0 mm to increase catalyst productivity. The catalyst effectiveness factor is reported to be at least 0.9 for 1.5 mm ammonia catalyst versus approximately 0.6 for 5.7 mm catalyst [12]. The catalyst volume can be reduced by 27 % if 2 mm catalyst particles are used for ammonia synthesis instead of 10 mm catalyst particles [3].

The design of the radial-flow reactor is more complicated than that of the axial-flow reactor because of several unique concerns. First, to achieve good conversion, the flow must be distributed uniformly along the length of the reactor. Second, catalyst settling over time will create a void at the top of a radial catalyst bed which, if not blocked, will allow flow bypassing. Depending on the mechanical properties of the catalyst and the frequency of reactor startup and shutdown, 5–15 % of extra catalyst is normally loaded in a radial-flow reactor to allow for settling. Third, compared with an axial-flow reactor, a radial-flow reactor requires additional void space (i.e., space not occupied by catalyst) for distributing flow along the length of the reactor. For some high-temperature reactions such as dehydrogenation of ethylbenzene to styrene, the thermal reactions taking place in the void are less selective than the catalytic reactions. To minimize the yield loss caused by the thermal reactions, commercial styrene reactors typically have a displacement cylinder in the center pipe to reduce the void volume. Finally, the perforated screens containing the catalyst must be designed to withstand the thermal stresses that develop with each startup and shutdown, especially during emergency shutdowns when large temperature differences may develop between the inner and outer screens.

Radial-flow catalyst beds may be susceptible to channeling under some low-flow conditions because pressure drops are too low to distribute the feed gas properly. A problem encountered in ammonia plants is the initial activation of the ammonia synthesis catalyst. Since at startup there is no heat of reaction, synthesis gas is heated in a startup heater and circulated through the ammonia reactor to reduce the catalyst. Unless the startup heater is very large, the gas velocity during startup is much lower than during normal operation. At such an extremely low velocity it is difficult to obtain uniform flow through the catalyst bed. Ideally, to avoid water damage to the catalyst that has already been reduced to free iron, catalyst reduction should proceed in a plug-flow manner, i.e., once the gas has picked up water vapor from catalyst reduction, it should pass only the catalyst that is still in oxidized state. Prereduced catalysts have been used in some radial-flow ammonia reactors because of concerns over flow channeling during startup [13].

2. Applications of Radial-Flow Reactors

The physical dimensions and operating conditions of several commercial radial-flow reactors are presented in Table 2. Illustrated in Figures 2,3, 4 and 5 are radial-flow reactor designs for ammonia synthesis, ethylbenzene dehydrogenation, and catalytic reforming.

Synthesis of ammonia from hydrogen and nitrogen (→ Ammonia) is an exothermic reaction, and is usually carried out at 400–500 °C and

Table 1. Comparison of axial- and radial-flow reactors for ammonia synthesis [8]

Operating variables	Small unit			Large unit	
	Axial	Radial I	Radial II	Axial	Radial
Space velocity, h^{-1}	25 000	25 000	30 000	25 000	25 000
Capacity, t/d	300	300	350	1 500	1 500
Total pressure drop, kg/cm^2	23	15.2	22.7	30	18.2
catalyst bed	8	0.2	0.2	12	0.2
other elements of reactor	3	3	4.5	4	4
condensing and piping	12	12	18	14	14
Power consumption kW·h/t NH_3	80	53	80	100	64

Table 2. Designs of commercial radial-flow reactors*

Reaction	Ammonia synthesis	Ethylbenzene dehydrogenation	Catalytic reforming 1st stage	Catalytic reforming, 2nd stage	Catalytic reforming, 3rd stage
Temperature, °C	380	589	505–535	505–535	505–535
Pressure, MPa	15.1	0.14 (1st bed)	3.5	3.5	3.5
Diameter of reactor shell, m	2.3	4.1			
Diameter of outer catalyst screen, m	2.1	3.4	2.2	3.0	4.0
Diameter of inner catalyst screen, m	0.4	1.5	0.7	0.7	0.7
Length of bed, m	13.5	8.1	5.0	5.8	6.1
Reference	13	23	24	24	24

* Note: The operating conditions are those available in the literature, but they may not represent the latest technologies.

14–21 MPa. The ammonia concentration in the reactor effluent is 12–22 %, which results in a large amount of the synthesis gases being recycled. Because the radial-flow reactor has a much smaller pressure drop, choosing a radial-flow reactor over an axial-flow reactor can reduce the size of the recycle gas compressor and allow use of smaller catalyst particles to improve the catalyst efficiency. Major ammonia technology licensors, Kellogg Brown & Root (KBR), Topsøe, Uhde, and Ammonia Casale, have developed their own proprietary radial-flow reactor designs [14 – 21]. Figure 2 shows KBR's Advanced Ammonia Process converter design. It has four catalyst beds in one shell, and three interstage coolers to remove the heat of reaction from the effluents of beds 1–3. In all four beds, the gas flows from outside to inside. In the upper three beds, the gas flows downward in the outer annulus and upward in the inner annulus (i.e., Π-shaped catalyst bed). After it comes out of the inner annulus, the gas turns downward to pass through the shell side of the interstage cooler. In the bottom bed, the gas flows downward in both the outer annulus and the center pipe (i.e., Z-shaped catalyst bed). Another radial-flow reactor design for ammonia synthesis, the Topsøe S-200 type converter, is illustrated in Figure 3. It has two catalyst beds and two internal heat exchangers: one at the bottom of the reactor and the other in the center core of the lower catalyst bed. The cold reactor feed first flows down the annulus between the reactor shell and the shell containing the catalyst bed. There is a large pressure difference but a small temperature difference across the reactor shell, and a small pressure difference but a large temperature difference across the catalyst bed shell. The feed is heated in the bottom heat exchanger by the effluent from the upper bed (the second catalyst bed). It then mixes with a cold bypass and with a second feed stream, which is heated in the interbed heat exchanger by the first bed effluent. The mixed feed flows from the annulus toward the center of the first catalyst bed. After passing through the shell side of the interbed exchanger, the first-bed effluent flows outward across the second catalyst bed, and then exits through an annulus and the bottom heat exchanger. The first catalyst bed has inward flow, while the second catalyst bed has outward flow. Both catalyst beds are Π-shaped.

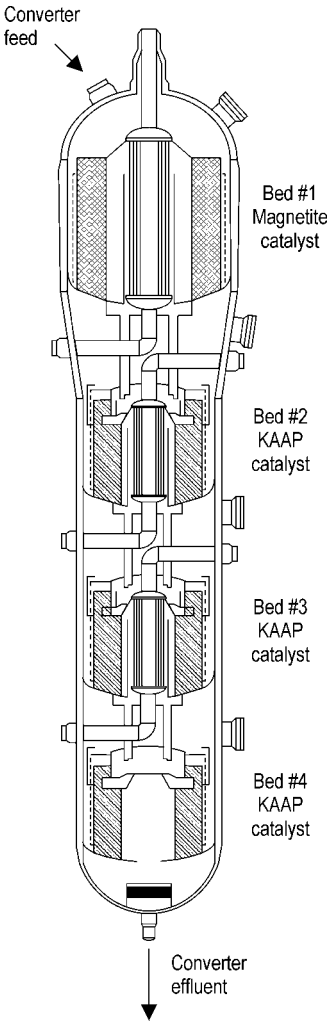

Figure 2. KAAP ammonia converter [14]

Dehydrogenation of ethylbenzene to styrene (→ Styrene) is an endothermic reaction. It is carried out with steam dilution at 540–650 °C and under low positive pressure or vacuum. The reaction is limited by thermodynamic equilibrium, and the reactant must be reheated between the catalyst beds in order to achieve 60–70 % per-pass conversion. Commercial styrene plants typically have two or three catalyst beds in series, each contained in individual vessels, with external or internal interstage reheaters. Both the conversion and selectivity are favored by lower pressure. Radial-flow reactors are used for ethylbenzene dehydrogenation because their smaller pressure drops allow the reactors to operate at a lower pressure than axial-flow reactors. Figure 4 shows a typical commercial styrene reactor design. In this design, the reactant flows from the central core toward the outer annulus. Since the reaction takes place under vacuum, the volume of the gas flow increases significantly as it moves across the catalyst bed. The outward flow arrangement provides a larger cross-sectional area toward the outlet of the catalyst bed as the gas volume increases. Many commercial styrene reactors have a displacement cylinder in the central core to reduce the void volume [22]. At temperatures above 540 °C, a small amount of ethylbenzene is dehydrogenated to styrene even in the absence of catalyst, but the thermal reaction has a lower selectivity to styrene than the catalytic reaction. Therefore it is desirable to minimize the void volume to improve the selectivity to styrene.

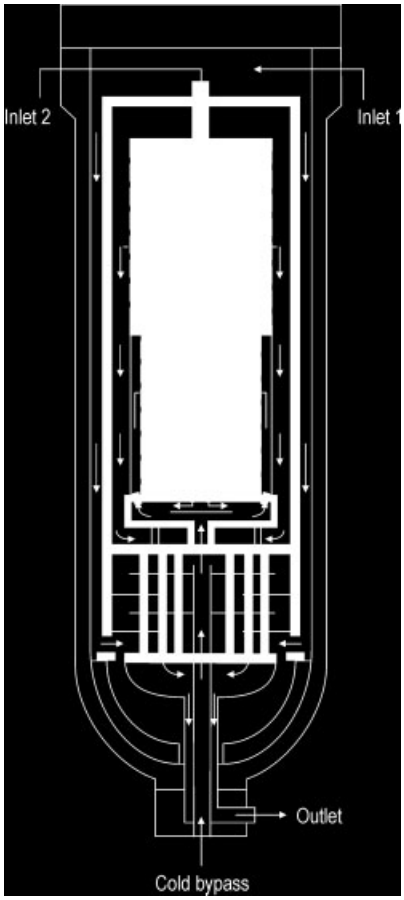

Figure 3. Topsøe S-200 type ammonia converter [17]

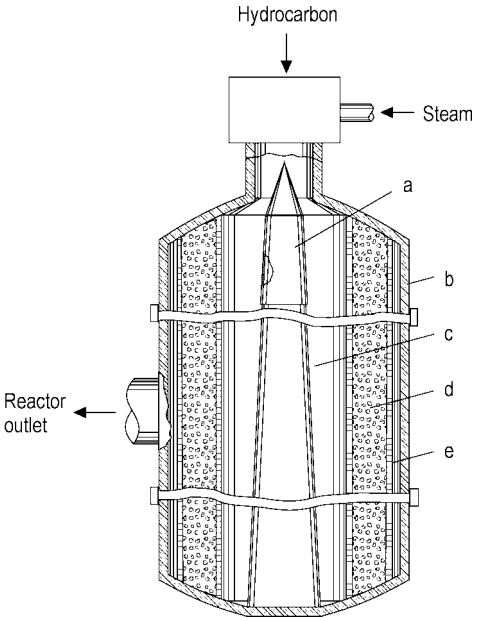

Figure 4. Radial-flow reactor for ethylbenzene dehydrogenation a) Displacement cylinder; b) Reactor shell; c) Void space; d) Catalyst; e) Catalyst basket [22]

Catalytic reforming is an important refining process for upgrading the octane number of naphtha for gasoline blending. It is also a major process for production of benzene, toluene, and xylenes for the petrochemical industry. In reforming, molecules are rearranged, or reformed, without changing their carbon numbers. Several reactions take place in catalytic reforming: dehydrogenation, isomerization, dehydrocyclization, and hydrocracking. The net effect of the reforming reactions is endothermic. The reaction temperature is in the range of 430–540 °C. The reaction pressure used to be above 2.8 MPa in order to control catalyst deactivation caused by coking. With more stable catalysts and continuous catalyst regeneration, the newer generation of reformers operates below 700 kPa. The reformer feed is a mixture of naphtha and hydrogen. Radial-flow reactors are used for catalytic reforming to reduce the pressure drop and thus the size of the hydrogen recycle compressor. Three or four reactors operate in series with external heaters in between; the reactors may stand alone or stack up [24 – 28]. Figure 5 shows a radial-flow, moving-bed reactor design developed by UOP for the CCR Platforming Process [27]. The catalyst moves slowly downward in

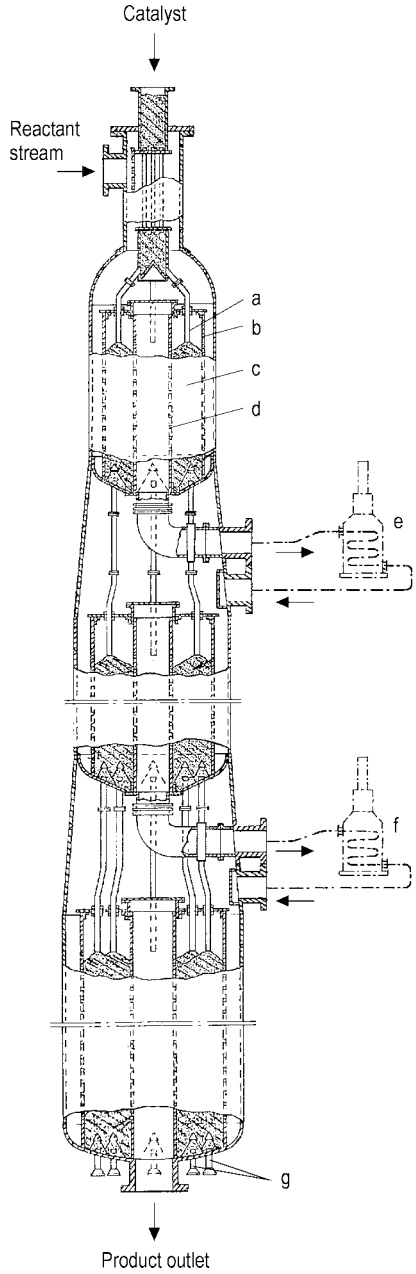

Figure 5. UOP radial-flow moving-bed reactor [27] a) Catalyst transfer pipe; b) Catalyst basket; c) Catalyst; d) Catalyst basket; e) Interstage heater; f) Interstage heater; g) Catalyst outlet [27]

the space contained between the outer and inner screens of the reactors, and through the pipes between the reactors. The catalyst coming out of the bottom reactor is lifted to the regenera-

tor and returned to the top reactor. The reactants flow from the outer annulus toward the center pipe, go out from the center pipe to the external heater, and return to the outer annulus of the reactor below. As the reactants move from one reactor to the other, the reaction rates become slower. To compensate for the slower reaction rates, the reactors become progressively larger, with the first reactor being the smallest and the last reactor the largest. In Figure 5, the effluent leaves the bottom of the center pipe. Since late 1980s, UOP has changed the design so that the effluent exits from the top of the center pipe, i.e., the catalyst bed configuration was changed from Z-shape to Π-shape [28]. As discussed in Chapter 4 this change improves the vapor distribution along the length of the catalyst bed.

UOP also developed similar radial-flow, moving-bed reactor designs for the dehydrogenation of propane to propylene (Oleflex Process) and the conversion of liquefied petroleum gas to aromatics (Cyclar Process).

3. Design of Commercial Radial-Flow Reactors

Many studies have been published in the literature on a wide range of subjects concerning radial-flow reactor design such as pressure drop [29], flow distribution [30 – 37], heat transfer in the radial direction [38], and effects of flow distribution and flow direction on conversion and selectivity [39 – 46]. The rigorous analyses and complex mathematical models presented in those publications are rarely used for commercial radial-flow reactor design. Reaction rates in most commercial radial-flow reactors are controlled by thermodynamic equilibrium. For these reactors, flow direction and axial and radial dispersions have small effects on conversion and selectivity. Pressure drops used for commercial radial-flow reactor design, in many cases, are empirical numbers derived from field data rather than from laboratory results. As explained in 5, the pressure drop increases over time as the reactor goes through startup and shutdown cycles.

Most problems with commercial radial-flow reactors are due to faulty mechanical design; they are not caused by lack of thermodynamic or kinetic data or inaccurate mathematical modeling. Many factors must be considered for the mechanical design of a commercial reactor: materials of construction, fabrication, transport, field erection, corrosion, thermal stresses, catalyst loading and unloading, and maintenance and repair of internal heat exchangers and catalyst basket screens. Case studies on the design and operation of commercial reactors have been published in *Ammonia Plant Safety*. Three subjects unique to radial-flow reactor design are discussed below: flow distribution along the length of the catalyst bed, catalyst basket design, and catalyst settling.

4. Flow Distribution

The principle governing the flow distribution in a radial-flow reactor is the same as that used for designs of sprayers for irrigation, pipe burners for gas fuel, diffusers for disposing of wastewater under sea, and air distributors for gas-turbine combustors. The basic unit of all these devices is a pipe with holes drilled through its wall (commonly referred to as a "manifold"): fluid may be collected from or distributed to the surroundings through the perforated wall. The fluid flow in a manifold may be described by a general form of the Bernoulli equation:

$$(1/\rho)\,\mathrm{d}P/\mathrm{d}Z+\kappa\,\mathrm{d}\left(U^2\right)/\mathrm{d}Z+\lambda\,(1/D)\left(U^2/2\right)=0 \tag{1}$$

where ρ is the density of the fluid, κ the kinetic energy coefficient (0.5 for pipe with a solid wall), and λ the friction factor. When a fluid moves along a manifold, its axial velocity changes as a result of the subtraction or addition of the side stream through the holes. If the friction loss is negligible, the pressure rises when the velocity decreases (loss in kinetic energy), and falls when the velocity increases (gain in kinetic energy). In contrast, friction always causes the pressure to decrease, regardless of whether the velocity increases or decreases.

In a manifold distributing fluid to the surroundings, if the influence of the kinetic energy is greater than that of the friction loss (e.g., a short manifold with a small length-to-diameter ratio), the pressure increases along the length of the manifold and more fluid is distributed toward the end of the manifold. Conversely, if the influence of the kinetic energy is smaller than

that of the friction loss (e.g., a long manifold with a large length-to-diameter ratio), the pressure decreases along the manifold and more fluid is distributed near the entrance of the manifold. Thus, one possible way to control the flow distribution is to vary the length-to-diameter ratio of a manifold, balancing the kinetic energy gain with the friction loss. Another way is to increase the resistance to the sidestream by decreasing the open area of the holes. As a rule of thumb, a uniform flow distribution can be obtained if the resistance to the sidestream is ten times or greater than the initial kinetic energy in the manifold. A third way to control the flow distribution is to vary the distribution of the holes. For example, placing more holes near the end of a slender manifold helps make the flow distribution more uniform.

In a manifold collecting fluid from the surroundings, the pressure always decreases toward the exit because of the friction loss and the gain in kinetic energy. Possible options to ensure a uniform flow distribution in a collecting manifold include increasing the resistance against the side stream and placing fewer holes near the exit.

In a radial-flow reactor the flow distribution is governed by the pressure profiles in the center pipe and the outer annulus and the resistance across the catalyst bed. Two separate equations can be written to describe the fluid flow in the center pipe and the outer annulus:

$$(1/\rho_{\mathrm{i}})\,\mathrm{d}P_{\mathrm{i}}/\mathrm{d}Z+\kappa_{\mathrm{i}}\,\mathrm{d}\left(U_{\mathrm{i}}^{2}\right)/\mathrm{d}Z+\lambda_{i}\,(1/D_{\mathrm{i}})\left(U_{\mathrm{i}}^{2}/2\right)=0 \tag{2}$$

$$(1/\rho_{\mathrm{o}})\,\mathrm{d}P_{\mathrm{o}}/\mathrm{d}Z+\kappa_{\mathrm{o}}\,\mathrm{d}\left(U_{\mathrm{o}}^{2}\right)/\mathrm{d}Z+\lambda_{\mathrm{o}}\,(1/D_{\mathrm{o}})\left(U_{\mathrm{o}}^{2}/2\right)=0 \tag{3}$$

where the subscript i refers to the center pipe, and o to the outer annulus. The difference between the pressures in the center pipe and the outer annulus is equal to the sum of the pressure drops across the inner screen, the catalyst bed, and the outer screen:

$$P_{\mathrm{i}}-P_{\mathrm{o}}=\Delta P_{\mathrm{i}}+\Delta P_{\mathrm{b}}+\Delta P_{\mathrm{o}}. \tag{4}$$

The pressure drops across the inner and outer screens can be calculated from the screen resistance coefficient ξ and the radial velocity, which varies from V_{i} to V_{o} across the catalyst bed:

$$\Delta P_{\mathrm{i}}=(\xi_{\mathrm{i}})\,(\rho/2)\,(V_{\mathrm{i}}/\Phi_{\mathrm{i}})^{2} \tag{5}$$

$$\Delta P_{\mathrm{o}}=(\xi_{\mathrm{o}})\,(\rho/2)\,(V_{\mathrm{o}}/\Phi_{\mathrm{o}})^{2} \tag{6}$$

where Φ_{i} and Φ_{o} are the void fractions of the inner and outer screens, respectively. The kinetic energy of the radial flow can be represented with the general form of the Bernoulli equation:

$$\mathrm{d}\,(P/\rho)\,/\mathrm{d}r+\mathrm{d}\left(V^{2}/2\varepsilon^{2}\right)+\lambda_{\mathrm{b}}\,(1/d_{\mathrm{p}})\left(V^{2}/2\right)(1-\varepsilon)\,/\varepsilon^{3}=0 \tag{7}$$

where ε is the void fraction of the catalyst bed, and λ_{b} is the friction factor of the catalyst bed and can be estimated from the Ergun equation:

$$\lambda_{\mathrm{b}}=150/Re_{\mathrm{p}}+1.75 \tag{8}$$

where Re_{p} is the Reynolds number based on the diameter of the catalyst particle and the local velocity. A computer model needs to be developed to solve all the above equations in order to calculate the flow distribution in a radial-flow reactor. More details are presented in [37].

Figure 6 illustrates how the reactor shape can affect the flow distribution in a radial-flow reactor. For the case that the friction loss controls the change in pressure, a Z-shaped reactor has a more uniform flow distribution than a Π-shaped reactor, because in a Z-shaped reactor the flows in the center pipe and the annulus move in the same direction, and so do the pressure changes. For the case that the kinetic energy predominates, a Π-shaped reactor has a more uniform flow distribution. In a Π-shaped reactor, the flows in the center pipe and the annulus move in opposite directions. If the pressure increases in the annulus because of deceleration, it will decrease in the center pipe because of acceleration. The reverse in flow direction offsets the pressure changes due to deceleration and acceleration, and gives a more uniform flow distribution in a Π-shaped reactor.

As shown in Figure 4, in a styrene reactor the reactant flows from the center pipe toward the annulus. The tapered displacement cylinder in the center pipe reduces the void volume to minimize thermal reactions. It also gradually reduces the cross-sectional area; thus, as the flow decreases along the center pipe, the velocity remains relatively constant, resulting in a flatter pressure profile and a more uniform flow distribution across the catalyst bed. The shape of the displacement cylinder is a variable that can be adjusted to improve the flow distribution in a radial-flow reactor.

To keep the diameter of the reactor shell within fabrication and transportation limits, an approach commonly used in scaling up radial-flow reactors is to increase the length without changing the diameters of the inner and outer screens and the reactor shell. However, as the capacity increases, the velocities in the center pipe and the annulus also increase. This results in steeper pressure profiles in the center pipe and the annulus and consequently a less uniform flow distribution across the catalyst bed. Certain reactor dimensions may need to be adjusted when the reactor design is modified for a greater capacity in order to maintain a uniform flow distribution.

5. Catalyst Basket Design

The screens of the catalyst basket in a radial-flow reactor may be formed with perforated plates or V-shaped "wires" welded circumferentially to supporting bars (commonly referred to by the trade name Johnson screens, see Figure 7). V-shaped or profile wires on Johnson screens are aligned vertically with the slots between them as small as 0.25 mm. This design provides a smooth surface which allows the catalyst to slide up and down without causing abrasion [47]. Perforated plates are expensive to fabricate for large reactors because the plates become too thick, and they have been gradually replaced by Johnson screens. To provide elasticity for thermal expansion, a layer of wire mesh is sometimes overlaid on the side of the perforated plate or the Johnson screen in contact with the catalyst. Outer screens of reformer catalyst baskets are commonly made of screens in a scallop shape [48, 60], which can be installed and repaired as individual pieces and thus more quickly and simply than a single large basket.

The Janssen equation, which was derived to describe the lateral pressure exerted on the wall of a silo [49], can also be applied to calculate the static force on the screens of a radial-flow reactor when the catalyst is initially loaded. As discussed below, the force on the screens becomes more complicated after the reactor goes through thermal cycling with startup and shutdown.

In a silo, the lateral pressure is transmitted to the wall through the interlocking force between the particles and the friction between the particles and the wall. The downward pressure on the bottom of a silo and the lateral pressure on the wall (the pressures in the axial and radial directions, respectively) can be described by the following equations:

$$P_z = (D\rho_b)\left(1-e^{(-4f'KZ/D)}\right) / (4f'K) \quad (9)$$

$$P_r = KP_z = (D\rho_b)\left(1-e^{(-4f'KZ/D)}\right) / (4f'), \quad (10)$$

where D is the diameter of the silo, ρ_b the bulk density of the solid particles, f' the wall friction factor, and Z the depth from the top. The ratio of lateral to downward pressure K is related to the internal friction between the particles; its value has been reported to be between 0.4 and 0.5 for a wide range of solid particles [50, 51].

Unlike in a liquid filled vessel, the downward pressure on the bottom of a silo is less than the weight of the granular material held in the silo. It levels off after the height reaches approximately three times the diameter. Additional weight beyond this height is supported by the friction at the wall. For an incremental depth ΔZ!

incremental friction force at the wall $= \Pi D \Delta Z P_r f\prime$, and

incremental solid weight $= \left(\Pi D^2/4\right) \Delta Z \rho_b$.

Equating the above two equations gives the asymptotic form of the lateral pressure

$$P_r = D\rho_b/4f\prime \quad (11)$$

The same equations developed for a silo can be used to calculate the radial pressure on the screens of a radial-flow reactor by replacing D with $D_o - D_i$. The friction factor depends on the shape of the catalyst and the surface condition of the screen, i.e., perforated wall, Johnson screen, or wire mesh; its value can be found only from experiments. Figure 8 illustrates the radial and tangential pressures on the inner and outer screens of an ammonia catalyst bed immediately after catalyst loading.

The Janssen equation only describes the force on the reactor screens when the catalyst is initially loaded. Additional forces are developed after each reactor startup and shutdown. For example, during the startup of an ammonia reactor, with the gas flowing from the annulus toward the center pipe, the outer screen is heated up first while the inner screen is still cold. As the outer screen is heated, it expands, the width between the inner and the outer screens increases, and the catalyst "settles" to fill the increased width. Because of the massive weight

of the catalyst, it takes several hours to heat up the catalyst and then the inner screen. After the catalyst has settled, the inner screen is finally heated, and it expands to push the catalyst against the outer screen. Thus, following each startup, more catalyst settles, the void above the catalyst bed increases, and greater force is exerted on the catalyst screens. During shutdown, the outer screen is cooled first; it contracts and compresses the catalyst against the inner screen. After both screens are cooled to ambient temperature, although the compressive force decreases, there is a net increase in the force on the screens compared with before the shutdown, because the outer screen contracts more than the inner screen. To avoid excessive stress on the screens, the inlet gas temperature should be increased or decreased slowly during startup or shutdown to minimize the temperature difference between the inner and the outer screens. Unfortunately, the inlet gas temperature often cannot be controlled during an emergency shutdown, and that is when most reactor screen failures happen.

A computer simulation [52] calculated the forces on the screens of an ammonia reactor based on the catalyst weight alone and those developed during normal startup (11 h) and fast startup (2.5 h). On a relative scale, if the force based on the catalyst weight is 1.0, the force resulting from a normal startup is 1.9, and from a fast startup 4.3.

The analysis described above can be applied to other radial-flow reactors. In the case of dehydrogenation of ethylbenzene to styrene, the reactant flows from the center pipe toward the annulus in a radial-flow reactor. During startup the catalyst is compressed first as the inner screen is heated and expands. It later settles when the outer screen is also heated and expands more than the inner screen. The reverse happens during shutdown: the catalyst first settles when the inner screen is cooled down and it is compressed later when the outer screen starts to cool down. This is an inherent feature of radial-flow reactor: differential thermal expansion between the inner and outer screens and the resulting catalyst settling and stress exerted on the screens after each startup and shutdown. The lifetime of catalyst screens depends to a large extent on the frequency of startup and shutdown that the reactor experiences.

Catalyst leakage from an ammonia reactor because the screens were twisted and torn apart, and the modified design of welding a perforated liner of Inconel to the inner side of the center pipe over the Johnson screens were described [53]. A reforming reactor was shut down after the skin thermocouple on the reactor outer shell recorded ca. 560 °C [54]. When the reactor was inspected, a hole was found on the screen and ca. 3.5 m of catalyst was found in the annulus. Because there was little flow at the bottom of the annulus to remove the heat generated from the hydrocracking and demethylation reactions, the temperature rose and triggered the shutdown. The catalyst in the bottom of the reactor was severely coked and fused into a solid mass.

6. Catalyst Settling and Flow Bypassing

Catalyst in a radial-flow reactor settles or slumps (i.e., drops in height) over time. A critical feature in the design of a radial-flow reactor is to prevent bypass at the top of the catalyst bed. Additional catalyst is loaded above the perforated section of the catalyst basket to compensate for catalyst settling and to seal the flow path above the perforation (Fig. 9).

The amount of catalyst needed for settling varies from process to process, and is determined from practical experience. As discussed in 5 , the catalyst settles as a result of the differential thermal expansion between the inner and outer screens of the catalyst basket. Thus, the thermal expansion coefficients of the materials used for the construction of the screens, the operating temperature, the rates of heating and cooling during startups and shutdowns, and the frequency of startups and shutdowns all affect catalyst settling. Loading procedure also affects the catalyst settling. Evenly spreading the catalyst to make the catalyst more densely packed and limiting the height of free fall to avoid breakup of catalyst particles can reduce the settling. Techniques such as dense loading and vibration have been used to pack catalyst more closely to reduce settling [12, 55]. Some catalysts contain volatile components and require special startup and shutdown procedures for conditioning. If proper startup and shutdown procedures are not followed, those components may leach out of the

reactor and cause the catalyst to shrink and drop. Catalyst shrinkage has been cited as a main reason why the radial-flow reactor is not as obvious a choice for methanol synthesis as for ammonia synthesis [56]. The methanol synthesis catalyst is mechanically much weaker than the ammonia synthesis catalyst, and it shrinks significantly during activation, which must take place in the reactor.

The flow direction across a radial-flow reactor can affect material selection for the catalyst screens and catalyst settling [57]. The design considered is for an ammonia synthesis reactor with the feed at 400 °C and the effluent at 510 °C. The diameter of the inner catalyst screen is approximately 90 cm, and the diameter of the outer screen 205 cm. For the outward-flow design, the inner screen is made of 2-1/4 Cr and 1 Mo and the outer screen 304 S.S. For the inward flow design, the inner screen is made of 304 S.S., and the outer screen 2-1/4 Cr and 1 Mo. Because it is rated for higher temperature service, 304 S.S. is selected for the screen where the effluent exits. Over the temperature range from 25 to 510 °C, the average thermal expansion coefficients of 304 S.S. and 2-1/4 Cr and 1 Mo are $18.4 \times 10^{-6}\ K^{-1}$ and $13.5 \times 10^{-6}\ K^{-1}$, respectively. The basket volume increases by 0.825 % for the inward-flow design and 1.966 % for the outward flow design, each time the reactor is heated from ambient temperature to its normal operating temperature. For this specific design, the outward-flow reactor will have more than twice as much catalyst settling as the inward-flow reactor, and will experience greater stress on the catalyst screens as it goes through the startup and shutdown cycles over its lifetime.

To monitor catalyst settling, a thermowell with multiple thermocouples can be inserted from the top to measure the temperature profile along the length of a radial-flow reactor. If the catalyst drops below the solid sections of the catalyst basket, the top thermocouple will measure a temperature much higher (endothermic reaction) or much lower (exothermic reaction) than the other thermocouples below.

The temperature profile also indicates if the reactant gas is distributed uniformly along the length of the catalyst bed. If it is, the temperature readings at different heights should be approximately the same.

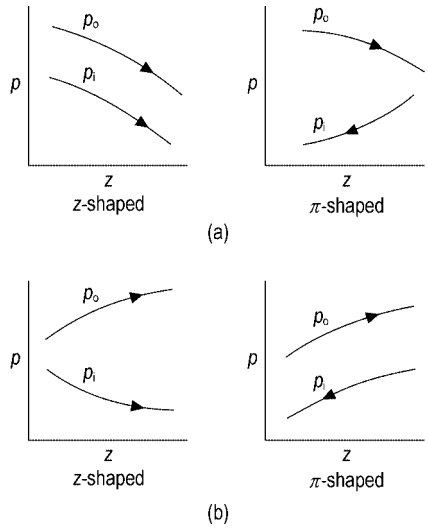

Figure 6. Effect of reactor geometry on flow distribution a) Friction loss dominates; b) Effect of deceleration and acceleration dominates

Special design is required to allow the thermowell to grow against the surrounding catalyst. The increase in length due to thermal expansion can be very significant: for example, a 15 m long, 304 S.S. reactor grows 14 cm when it is heated from 25 to 540 °C. Some companies such as Gayesco specialize in providing thermowells and thermocouples for this type of reactor service.

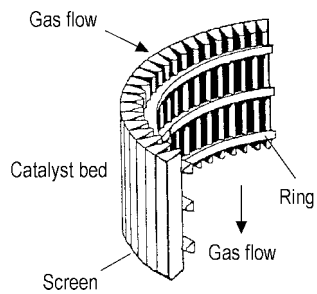

Figure 7. Catalyst basket screen [53]

As illustrated in Figure 9, some flow may channel through the catalyst above the perforated section of the basket. In the extreme case, the upward vapor flow fluidizes the catalyst, and the catalyst fines generated from fluidization penetrate into the catalyst bed and the downstream screens, resulting in a significantly

greater pressure drop across the reactor. The catalyst loaded above the perforated section should be sufficient to allow for settling and to prevent channeling. In some radial-flow reactors, a baffle plate is mounted above the perforation to further restrict the channeling.

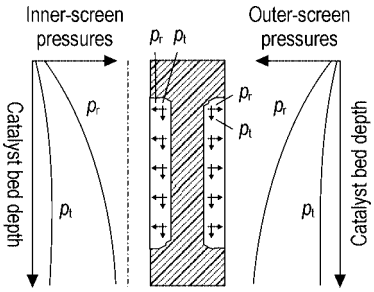

Figure 8. Radial (P_r) and tangential (P_t) pressures on the inner and outer screens of an ammonia converter after catalyst loading [52]

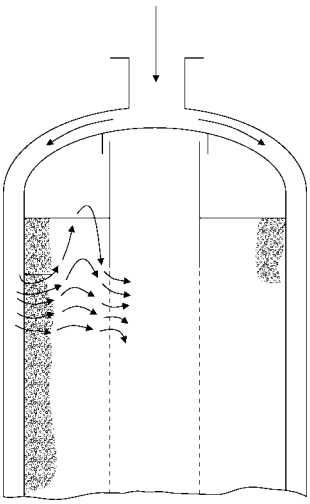

Figure 9. Top section of a radial-flow reactor

Ref. [14] describes experiences at two 1850 t/d ammonia plants that were forced to shut down because of excessively high pressure drops across the reactors, and how the designs of the reactors were modified to solve the pressure drop problems. Both plants use Kellog Brown & Root's Advanced Ammonia Process (KAAP) technology, which has four radial catalyst beds in the reactors. The first bed, packed with iron catalyst, is of the conventional design with no reactant flowing through the settling zone (called freeboard zone in [14]). The lower three beds, which are packed with high-activity, precious metal based catalyst, have a KAAP patented design that is composed of three major sections: freeboard, chute, and main bed (see Fig. 10). The reactant is directed to pass through the freeboard zone first before it reaches the main bed. The chute is the conduit whereby the catalyst moves from the freeboard zone into the main bed as settling occurs. This unique design allows 100 % utilization of all catalyst loaded. The freeboard section is subdivided into two parts: a freeboard catalyst zone and a freeboard bypass zone. A portion of the feed gas passes through the catalyst zone, mixes with the balance of the feed gas from the bypass zone, and the combined gas stream then flows to the main bed. Initially a screen was placed between the freeboard catalyst zone and the freeboard bypass zone. After some investigations with CFD modeling and a cold mock-up test unit, it was discovered that there was enough upflow from the catalyst zone to the bypass zone to fluidize the catalyst, and the catalyst particles pushed through the gaps between the screen panels plugged the inlet screens of the main bed, causing the pressure drop to increase. At one plant, the screen between the freeboard bypass zone and the freeboard catalyst zone was replaced with a solid plate. At the other plant, the opening at the outlet of the freeboard bypass zone was blanked, forcing the gas to flow downward through the catalyst zone. Both modifications stopped the increase in reactor pressure drop.

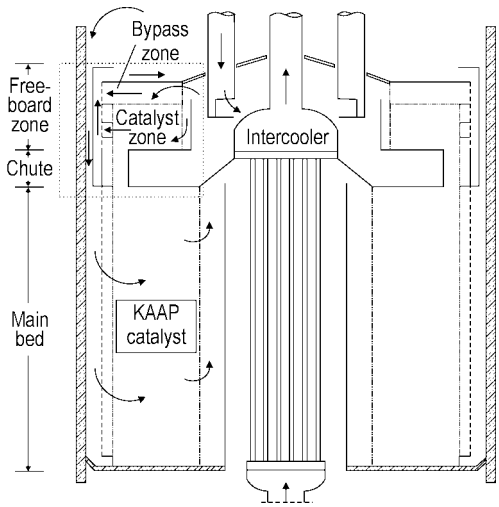

Figure 10. KAAP reactor catalyst bed [14]

To utilize the catalyst reserved for settling (15 % of the total loading), Axens, a subsidiary of IFP, developed the Texicap design: a flexible flow guide that settles with the catalyst to prevent bypass across the gap above the catalyst bed [58]. For a similar purpose, Ammonia Casale developed the axial-radial flow bed design, which allows about 90 % of the feed gas to pass through the catalyst bed in a radial direction and the balance to pass down through a top layer of catalyst in an axial direction (see Fig. 11). In addition to ammonia synthesis, Ammonia Casale has applied this design to the high-temperature CO shift reaction for the revamp of an ammonia plant [59]. It claimed that the water droplets leaking from an upstream steam generator forms a layer of cake preferentially on top of the axial bed, shifting more gas flow toward the radial catalyst bed and minimizing the effect on the total reactor pressure drop.

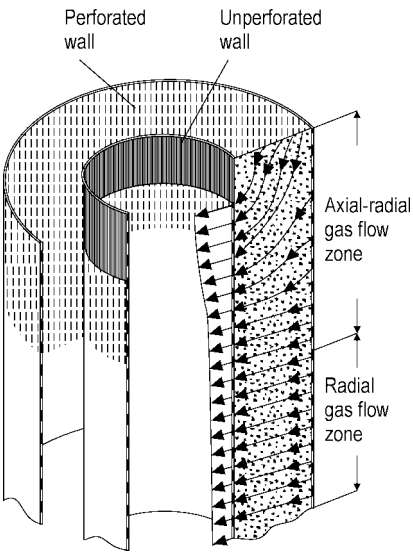

Figure 11. Casale axial-radial catalyst bed [20]

7. Acknowledgments

The author first started his study of radial-flow packed-bed reactors while he was a graduate student at the University of Washington. He would like to express his appreciations to Professor Bruce A. Finlayson for suggesting radial-flow reactor as the subject of his Ph.D. thesis and to the Department of Chemical Engineering at the University of Washington for the financial aids that supported his graduate study. The author would also like to thank Sanjeev Ram of ABB Lummus for his valuable comments on this paper.

8. References

1. "Radial Flow Ammonia Converter—New Ammonia Synthesis Design," *Nitrogen* **31** (September 1964) 22.
2. "Giant Ammonia Converters", *Nitrogen* **75** (January/February 1972) 33.
3. "Cutting Big Plants Down to Size", *Chemical Week* (February 26 1966) 97.
4. "Giant Ammonia Plant Hikes U.S. Output", *Oil Gas J.* (October 23 1967) 52.
5. M. G. Slin'ko, V. S. Beskov: "Basic Directions in the Development and Design of High-Capacity Reactors", *Sov. Chem Ind.* **11** (1979) no. 4, 251.
6. T. V. Dodogorakaya, S. P. Sergeev, R. S. Veranyan: "High-Unit Capacity Catalytic Reactors for Ammonia Plants", *Sov. Chem. Ind.* **11** (1979) no. 3, 187.
7. M. B. Aizenbud et al.: "The Development and Introduction of Ammonia Synthesis Reactors with Low Hydraulic Resistance", *Sov. Chem. Ind.* **9** (1977) no. 8, 624.
8. V. N. Gaganova, S. S. Lachinov, N. S. Torocheshnikov: "Prospects for the Use of Radial Ammonia Synthesis Columns", *Moskovskii Khimiko-Tekhnologich-eskii Institut imeni D. I. Mendeleeva Trudy* **65** (1970) 132.
9. R. Hakmann et al.: "Revamping Ammonia Converters", *Ammonia Plant Safety* **26** (1986) 44.
10. R. L. Newland, J. D. Pierce, D. M. Borzik: "Revamped Ammonia Plant Utilizing the Topsoe Series-200 Radial Flow Converter", *Ammonia Plant Safety* **28** (1988) 53.
11. K. A. Clayton, N. Shannahan, B. Wallace: "Retrofitting the Ammonia Syngas Converter for Energy Efficiency", *Ammonia Safety* **31** (1991) 212.
12. A. Nielsen: "Radial Flow Ammonia Converter", *Ammonia Plant Safety* **14** (1971) 46.
13. Haldor Topsøe, US 3372988, 1964 (H. J. Hansen).

14. K. Blanchard, S. Noe, E. Plaxco: "KAAP Converters: Inspection, Remediation, and Modification", *Ammonia Plant Safety* **41** (2001) 284.
15. T. A. Czuppon et al.: "First Commercial Application of the Kellogg Advanced Ammonia Process", *Ammonia Plant Safety* **34** (1994) 236.
16. T. A. Czuppon, S. A. Knez, R. B. Strait: "Commercial Experience with KAAP and KRES", *Ammonia Plant Safety* **37** (1997) 34.
17. Haldor Topsoe brochure.
18. D. Novacek, R. Horsey, R. Newland: "In-Situ Series-200 Modification of Ammonia Converters", *Ammonia Plant Safety* **30** (1990) 161.
19. L. Sutherland, B. Wallace: "Improvements to CFI's Ammonia Converters Utilizing Casale's Internals", *Ammonia Plant Safety* **29** (1989) 167.
20. U. Zardi: "Review these Developments in Ammonia and Methanol Reactors", *Hydrocarbon Processing* (August 1982) 129.
21. N. Shannahan: "Ammonia Loop Modernization", *Hydrocarbon Processing* (January 1989) 60.
22. Dow Chemical, US 4132743, 1978 (W. M. Castor, B. S. Taylor).
23. H. F. Rase, *Fixed-Bed Reactor Design and Diagnostics*, Butterworths Publishers, Boston 1990, p. 276.
24. T. Z. Khuramshin et al.: "Experience in Operating Catalytic Reforming Reactor with Radial Movement of Feedstock", *Chem. Technol. Fuels Oils* (1974) no. 2 134.
25. A. M. Sycheva et al.: "Experience in Operating Catalytic Reforming Reactor with Radial Feed Input and Thin Bed of Catalyst", *Chem. Technol. Fuels Oils* (1976) no. 1, 43.
26. E. L. Lobanov, Yu. A. Skipin: "Increasing the Operating Efficiency of Radial Reactors in Reforming", *Chem. Technol. Fuels Oils* (1986) no. 6, 275.
27. UOP, US 3706536, 1971 (A. R. Greenwood et al.).
28. N. L. Gilsdorf et al.: "Ultra-Low-Pressure Reforming Process" in J. J. McKetta (ed.): *Encyclopedia of Chemical Processing and Design*, Vol. **47**, Marcel Dekker, New York 1994, p. 94.
29. A. M. Scheva et al.: "The Resistance of a Bed of Packing During Radially Directed Flow", *Int. Chem. Eng.* **10** (1970) no. 1, 66.
30. E. A. Reizes: "Hydraulic Calculation for Radial Contact Apparatus", *Theor. Found. Chem. Eng.* **1** (1967) no. 3, 308.
31. I. D. Rokos et al.: "Calculating the Uniformity of Distribution of a Stream into a Catalyst Layer in the Case of a Reactor with Radial Charge Feed", *Chem. Technol. Fuels Oils* (1969) No.11/12, 824.
32. I. E. Idel'chik: "Equations for Calculating the Flow Distribution in Contact, Filtering, and other Devices, and Z-shaped Collectors", *Theor. Found. of Chem. Eng.* **4** (1970) no. 2, 234.
33. V. S. Genkin, V. V. Dil'man, S. P. Sergeev: "Distribution of Gas Flow Along Height of Catalyst Bed in Radial Catalysis Units", *Sov. Chem. Ind.* **4** (1972) no. 2, 121.
34. A. M. Sycheva et al.: "Effect of Ratio of Catalyst Bed Height to Vessel Diameter on Distribution of Flow in Reactors with Radial Inlet", *Chem. Technol. Fuels Oils* (1973) no. 5, 372.
35. V. V. Dilman et al.: "Calculation of the Hydraulic Resistance of Radial Contact Equipment", *Theor. Found. Chem. Eng.* **10** (1976) no. 4, 560.
36. R. Z. Adinberg et al.: "Hydraulics of an Industrial Radial Reactor for Conversion of Carbon Monoxide", *Sov. Chem. Ind.* **8** (1976) no. 6, 467.
37. C. H. Li, Flow Distribution in Radial Flow Packed Bed Reactors, Ph.D. Thesis, University of Washington, Seattle 1978.
38. J. Votruba, V. Hlavacek: "An Experimental Study of Radial Heat Transfer in Catalytic Radial-Flow Reactors", *Chem Eng. J.* (1972) no. 4, 91.
39. Y. A. Raskin et al.: "Mathematical Model and Calculation Algorithm for Radial Adiabatic Reactors", *Theor. Found. Chem. Eng.* **2** (1968) no. 2, 186.
40. V. A. Rastaturin: "The Choice of the Optimum Construction for an Adiabatic Reactor and its Main Dimensions", *Int. Chem. Eng.* **11** (1971) no. 1, 44.
41. J. M. Calo: "Cell Model Studies of Radial Flow, Fixed Bed Reactors", *ACS Symposium Series* **65** (1977) 550.
42. P. R. Ponzi, L. A. Kaye: "Effect of Flow Maldistribution on Conversion and Selectivity in Radial Flow Fixed-Bed Reactors", *AIChE J.* **25** (1979) no. 1, 100.
43. V. Balakotaiah, D. Luss: "Effect of Flow Distribution on Conversion in Isothermal Radial Flow Fixed-Bed Reactors", *AIChE J.* **27** (1981) no. 3, 442.

44. H. C. Chang, J. M. Calo: "An Analysis of Radial Flow Packed Bed Reactors – How Are They Different?", *ACS Symp. Ser.* **168** (1981) 305.
45. H. C. Chang, M. Saucier, J. M. Calo: "Design Criterion for Radial Flow, Fixed-Bed Reactors", *AIChE J.* **29** (1983) no. 6, 1039.
46. C. S. Yoo, A. G. Dixon, "Modeling and Simulation of a Radial Flow Fixed-Bed Reactor", paper presented at the 1985 Annual AIChE Meeting (Chicago Nov. 10–15, 1985).
47. UOP, US 4276265, 1979 (G. A. Gillespie)
48. D. M. Little, *Catalytic Reforming*, Pennwell Publishing Co., Tulsa Oklahoma 1985, pp. 35–37.
49. W. L. McCabe, J. C. Smith, *Unit Operations of Chemical Engineering*, 3rd ed., McGraw-Hill, New York 1976, pp. 812–814.
50. A. W. Jenike, J. R. Johanson, J. W. Carson: "Bin Loads – Part 3: Mass Flow Bins", *J. Eng. Ind. Trans. ASME* (Feb. 1973) 6.
51. W. Reisner, M. V. Eisenhart Rothe, *Bins and Bunkers for Handling Bulk Materials – Practical Design and Techniques*, Trans. Tech. Publications, Cleveland Ohio 1971, p. 47.
52. S. G. Thomsen: "Impact of Catalyst Forces on the Ammonia Converter Internals", *Ammonia Plant Safety* **40** (2000) 82.
53. M. A. Khan, J. J. Rasmussen: "Experience of Ammonia Converter with Catalyst Leakage and Repairs", *Ammonia Plant Safety* **40** (2000) 33.
54. D. M. Little, *Catalytic Reforming*, Pennwell Publishing Co., Tulsa Oklahoma 1985, pp. 138–139.
55. E. Dessen, G. Ryntveit, Y. Haquet: "Dense Loading of Catalysts in Ammonia Plants", *Ammonia Plant Safety* **33** (1993) 18.
56. I. B. Dybkjaer: "Design of Ammonia and Methanol Synthesis Reactors" in H. I. de Lasa (ed.): *Chemical Reactor Design and Technology*, NATO ASI Series, Martinus Nijhoff Publishers, 1986, pp. 808–809.
57. Lummus Crest, US 4461751, 1983 (J. V. Albano, G. Friedman).
58. J. Ross: "Balancing Octane and Hydrogen with ULS Fuels", *Petroleum Technol.* (2004) 49.
59. I. Wood, E. Filippi: "HTS Converter Revamping at Ammonia Plant: Case History", *Ammonia Plant Safety* **40** (2000) 164.
60. UOP, US 6224838, 1999 (G. A. Schulz, J. E. Burgard).